Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Anastassia Ugrjumova

Mudelipõhise klasteranalüüsi ja K-medoidide meetodi võrdlemine kvalitatiivsete tunnustega andmete klasterdamisel

Matemaatika ja statistika õppekava
Matemaatilise statistika eriala
Magistritöö (30 EAP)

Juhendaja: Kristi Kuljus

Tartu 2020

Mudelipõhise klasteranalüüsi ja K-medoidide meetodi võrdlemine kvalitatiivsete tunnustega andmete klasterdamisel

Magistritöö

Anastassia Ugrjumova

Lühikokkuvõte. Magistritöö eesmärk on võrrelda kaht erinevat klasteranalüüsi meetodit, kus üks on mudelipõhine ja teine põhineb vaatlustevahelistel kaugustel. Täpsemalt, võrreldakse mudelipõhist klasteranalüüsi ja K-medoidide meetodit kvalitatiivsete tunnuste korral. K-medoidide meetodi rakendamiseks kasutatakse PAM-algoritmi *(partitioning around medoids*). Mudelipõhise klasteranalüüsi puhul on vaatlused kirjeldatud segujaotuse abil, samal ajal PAM-algoritm põhineb erinevusmõõtudel. Viiakse läbi simulatsioonid erinevate klastrite kattuvusmäärade korral ja uuritakse mõlema klasterdusmeetodi käitumist erinevate kattuvuste korral. Et tulemusi analüüsida, kasutatakse kohandatud Randi indeksit ja keskmise silueti laiuse kriteeriumit.

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

Märksõnad: keskmise silueti laiuse kriteerium, klasteranalüüs, klastrid, kohandatud Randi indeks, mudelid, PAM-algoritm, tõenäosusjaotused, simulatsioon, R (programmeerimiskeel).

Comparison of Model-Based Clustering and K-medoids method for Clustering Categorical Data

Master's thesis

Anastassia Ugrjumova

Abstract. The aim of this master's thesis is to compare two different cluster analysis methods, where one is model-based and another one is a distance-based method. Specifically, model-based approach and K-medoids method are compared for categorical data. For applying the K-medoids method, PAM algorithm (partitioning around medoids) is used. For model-based clustering observations are described by a mixture distribution, whereas PAM algorithm uses dissimilarity measure. Simulations with different cluster overlapping are carried out and performance of both clustering methods is studied under different overlapping parameters. To analyse the results, the adjusted Rand index and the average silhouette width are used.

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

Keywords: adjusted Rand index, average silhouette width, cluster analysis, clusters, models, PAM algorithm, probability distributions, simulation, R (programming language).

Sisukord

Sissejuhatus

Antud magistritöö eesmärk on viia läbi klasteranalüüs K-medoidide meetodi ja mudelipõhise klasteranalüüsi abil kvalitatiivsete tunnuste jaoks ja võrrelda saadud tulemusi. Klasteranalüüsi eesmärgiks on grupeerida andmed ehk leida klastrid nii, et sama grupi ehk klastri objektid oleksid võimalikult sarnased ja erinevate klastrite objektid võimalikult erinevad. Sellist analüüsi on võimalik teostada meetoditega, mis võivad põhineda nii vaatlustevahelisel kaugusel kui ka vaatlusi kirjeldaval tõenäosusjaotusel, seega „sarnasuse“ ja „erinevuse“ mõisted on nende meetodite puhul erinevad. Kaugusel põhinevaks meetodiks on antud töö raames K-medoidide meetod, mille rakendamiseks kasutatakse PAM-algoritmi (*partitioning around medoids*). PAM-algoritmi korral mõõdetakse objektidevahelist erinevust ja öeldakse, et objektid on sarnased, kui klastrisisesed objektidevahelised erinevused on väikesed. Samal ajal mudelipõhise klasteranalüüsi korral on vaatlused kirjeldatud parameetrilise tõenäosusjaotuse abil ja klastrid on defineeritud segujaotuse komponentide kaudu.

Käesoleva töö idee tuleneb artiklist Anderlucci ja Hennig (2014), kus võrreldakse mudelipõhist klasteranalüüsi ja PAM-algoritmi, uurides klasterdamist väiksema ja suurema klastrite kattuvuse korral ning vaadeldes erinevaid ja võrdseid segujaotuse komponentide kaalusid ning erinevat arvu kvalitatiivsete tunnuste võimalikke väärtusi. Antud töös aga genereeritakse kvalitatiivsete tunnustega andmestikud, milleks kasutatakse etteantud klastrite kattuvusi ning erinevaid ja võrdseid segujaotuse komponentide kaalusid. Kui klastrite kattuvus on „suur“, siis on arvatavasti raske klastreid eraldada, kuid mida tähendab „suur“ klastrite kattuvus ei ole ette teada. Uuritakse, millist kattuvust saab „suureks“ nimetada ja milline meetod saab suurema kattuvuse korral paremini klastreid eraldada. Klasteranalüüsi läbi viimiseks kasutatakse tarkvara R lisapakettide funktsioone.

Töö esimeses peatükis defineeritakse erinevusmõõdud, mis sobivad kvalitatiivsete tunnuste klasterdamiseks. Antud töös vaadeldakse lihtsat sarnasuskoefitsienti ja Jaccardi koefitsienti ning selgitatakse välja, mis tingimustega on seotud ühe või teise koefitsiendi valik. Täpsemalt, uuritakse tunnuste sümmeetrilisust ja asümmeetrilisust ning sellega kaasnevaid eeldusi ja puudusi sarnasusmõõdu valikul. Sarnasusmõõdu illustreerimiseks tuuakse kaks näidet, kus lisaks on ühe näite eesmärk rõhutada, kui oluline on teha kindlaks, kas tunnus on sümmeetriline või asümmeetriline.

Töö teises osas käsitletakse K-medoidide meetodit ning vaadatakse detailselt läbi, kuidas teostatakse klasterdamist K-medoidide meetodil PAM-algoritmi abil. Samuti võrreldakse K-

medoidide meetodit K-keskmiste meetodiga, mis on väga levinud klasteranalüüsi meetodite hulgas. Tuuakse kaks näidet, millest üks on PAM-algoritmi illustreerimiseks ja teine on K-medoidide ja K-keskmiste meetodite klasteranalüüsi tulemuste võrdlemiseks binaarsete tunnuste jaoks.

Töö kolmandas peatükis kirjeldatakse mudelipõhist klasteranalüüsi: defineeritakse segujaotus kvalitatiivsete tunnuste jaoks, vaadeldakse segumudelite parameetrite hindamist EM-algoritmi abil, tuuakse välja hinnatavate segumudelite klassid erinevate kitsenduste korral. Samuti selgitatakse välja, milliseid kriteeriume kasutatakse mudelipõhise klasteranalüüsi parima mudeli valimisel. Osutub, et selleks on integreeritud klassifitseerimistõepära kriteerium ICL_{bic}, mis põhineb Bayesi informatsioonikriteeriumil.

1 Erinevusmõõdud kvalitatiivsete tunnuste korral

Kvantitatiivsete tunnuste korral on vaatlustevahelise erinevuse mõõtmiseks loomulik kasutada eukleidilist ja Manhattani kaugust. Olgu etteantud vaatlused $\boldsymbol{x_i}$ ja $\boldsymbol{x_j}$, mille omavahelist kaugust soovitakse mõõta. Defineerime Minkowski kauguse:

$$D(\boldsymbol{x_i}, \boldsymbol{x_j}) = \left(\sum_{l=1}^{d} |x_{il} - x_{jl}|^p \right)^{1/p}.$$

Kui $p = 2$, siis on tegemist eukleidilise kaugusega, ja kui $p = 1$, siis Manhattani kaugusega. Kvalitatiivsete tunnuste korral kasutatakse aga erinevusmõõtusid (*dissimilarity measure*). Tihti on erinevusmõõdud defineeritud läbi sarnasusmõõtude. Sellisel juhul on erinevusmõõt D defineeritud kui $D = 1 - S$, kus S on etteantud sarnasusmõõt, mille väärtused on lõigus $[0,1]$. Järgmiste alapeatükkide allikana kasutatakse raamatu Xu ja Wunsch (2008) teist peatükki, kui ei ole viidatud teisiti.

1.1 Sarnasusmõõdud binaarsete tunnuste korral

Olgu vaatlused $\boldsymbol{x_i}$ ja $\boldsymbol{x_j}$ kirjeldatud p binaarse tunnuse abil ja olgu vaatluste võimalikud väärtused 0 ja 1. Sarnasusmõõtude arvutamiseks vaadeldakse iga tunnuse korral vaatluste võimalike väärtuste paare ja samad paarid summeeritakse üle tunnuste kokku. Kui mingite tunnuste korral on mõlemal vaatlusel samaaegselt väärtus 1, st $\boldsymbol{x_{ik}} = 1$ ja $\boldsymbol{x_{jk}} = 1$ mingi tunnuse k korral, siis tähistatakse kõigi selliste paaride arvu n_{11}. Paaride arvu, mille korral mõlema vaatluse väärtus on 0, tähistatakse n_{00}. Kui mingite tunnuste korral tekivad paarid erinevate väärtustega, siis tähistatakse nende paaride arvu n_{10} (kui näiteks $\boldsymbol{x_{ik}} = 1$, $\boldsymbol{x_{jk}} = 0$) ja n_{01} (kui näiteks $\boldsymbol{x_{ik}} = 0$, $\boldsymbol{x_{jk}} = 1$), vaata tabelit 1.

Tabel 1. Objektide $\boldsymbol{x_i}$ ja $\boldsymbol{x_j}$ tunnuste väärtuste paaride sagedustabel

		Objekt $\boldsymbol{x_j}$: 1	Objekt $\boldsymbol{x_j}$: 0	
Objekt $\boldsymbol{x_i}$	1	n_{11}	n_{10}	$n_{11} + n_{10}$
	0	n_{01}	n_{00}	$n_{01} + n_{00}$
		$n_{11} + n_{01}$	$n_{10} + n_{00}$	p

Binaarseid tunnuseid saab jaotada kahte klassi järgmiselt: sümmeetrilised ja asümmeetrilised. Sümmeetrilise tunnuse korral on mõlemad tunnuse väärtused samaväärsed. Sümmeetriliseks tunnuseks on näiteks „sugu“ võimalike väärtustega „mees“ ja „naine“. Sellise tunnuse korral võib mõlemaid väärtusi tähistada nii 0 kui ka 1-ga, sest et need on võrdselt olulised sündmused. Järelikult n_{11} ja n_{00} on sama tähtsusega. Asümmeetriliseks tunnuseks peetakse tunnust, mille võimalikud väärtused omavad erinevat tähtsust. Tavaliselt olulisemat väärtust tähistatakse 1-ga. Kui binaarse tunnuse väärtused on „silmad on sinised“=1 ja „silmad ei ole sinised“=0, siis tunnus on asümmeetriline, sest et tähtsust omavad ainult sinised silmad. Kui silmad ei ole sinised, siis need võivad olla nii pruunid, rohelised kui ka hallid, st kui tunnuse väärtus on mõlema vaatluse korral 0, siis ei saa väita, et vaatlused on sarnased. Sellisel juhul n_{11} näitaks oluliste paaride arvu ja n_{00} ei oleks antud situatsioonis sama tähtsusega. Paneme tähele, et tunnuste sümmeetrilisus ja asümmeetrilisus sõltub vaadeldavast kontekstist.

Mõnikord on asümmeetriliste tunnuste kasutamine vajalik näiteks meditsiinivaldkonnas, kui soovitakse uurida mõnda haruldast juhtumit. Raamatu Kaufman ja Rousseeuw (1990) esimeses peatükis tuuakse näide veretüübi AB kohta, mida loetakse haruldaseks. Sellise tunnuse väärtuste „negatiivne“=0 ja „positiivne“=1 korral oleks andmestik täis väärtusi 0 ja objektide paarid, mille korral on mõlemad väärtused 0, ei näitaks kahe indiviidi sarnasust. Järelikult, kui veretüüp AB on negatiivne, siis ei saa väita, et indiviididel on midagi ühist. Seega asümmeetrilisuse ignoreerimine võib viia valede järeldusteni ning sümmeetriliste tunnuste jaoks mõeldud erinevusmõõdu kasutamine ei ole sobilik.

Sümmeetriliste tunnuste korral kasutatakse sarnasusmõõte, mille korral tähistuste 0 ja 1 vahetamine ei muuda tulemust ehk kõiki tunnuseid peetakse sümmeetrilisteks. Üks selline sarnasusmõõt on *lihtne sarnasuskoefitsient*.

Definitsioon 1. *Lihtne sarnasuskoefitsient näitab vaatluste keskmist sarnasuste arvu ja on defineeritud kujul*

$$S(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{n_{11} + n_{00}}{n_{11} + n_{00} + n_{10} + n_{01}} = \frac{n_{11} + n_{00}}{p}.$$

Antud sarnasusmõõdu põhjal defineeritud erinevusmõõt on Hammingu erinevus, see näitab vaatluste keskmist erinevuste arvu:

$$D(\boldsymbol{x_i}, \boldsymbol{x_j}) = 1 - S(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{n_{10} + n_{01}}{p}.$$

Asümmeetriliste tunnuste korral vaadeldakse sarnasusmõõte, mis ei võta arvesse paaride arvu n_{00}. Nende mõõtude defineerimisel kasutatakse ainult olulisemate paaride arvu n_{11}.

Definitsioon 2. *Jaccardi koefitsient on sarnasusmõõt, mis võtab arvesse ainult paaride arvu* n_{11} *ja see on defineeritud järgmiselt:*

$$S(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}.$$

Antud sarnasusmõõdule vastav erinevusmõõt on

$$D(\boldsymbol{x_i}, \boldsymbol{x_j}) = 1 - S(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{n_{10} + n_{01}}{n_{11} + n_{10} + n_{01}}.$$

Paneme tähele, et kui tegu on nii sümmeetriliste kui ka asümmeetriliste tunnustega, siis tuleb seda arvesse võta ja võib-olla on sel juhul sobilikum kasutada segatüüpi tunnuste jaoks mõeldud Gower'i erinevusmõõtu.

Näide 1. Vaatleme kolme binaarset tunnust, milleks on „sugu" väärtustega „mees"=0 ja „naine"=1, „vallaline" väärtustega „ei"=0 ja „jah"=1 ning „taimetoitlane" väärtustega „olen taimetoitlane"=0 ja „ei ole taimetoitlane"=1. Olgu nende tunnuste väärtused indiviidi $\boldsymbol{x_1}$ korral 0, 1, 0 ja indiviidi $\boldsymbol{x_2}$ korral 1, 0, 0. Paneme tähele, et $n_{10} = 1$, $n_{01} = 1$ ja $n_{00} = 1$. Seega antud vaatluste Hammingu erinevus on 2/3 (kui eeldatakse, et kõik tunnused on sümmeetrilised) ja Jaccardi erinevusmõõt on 1 (kui eeldatakse, et kõik tunnused on asümmeetrilised). Saadud erinevusmõõtude erinevad väärtused võivad viia erineva klasterduseni.

Näiteks raamatu Kaufman ja Rousseeuw (1990) esimeses peatükis vaadeldakse näidet, kus rõhutatakse erinevusmõõdu valiku olulisust. Olgu antud neli vaatlust, mis on kirjeldatud kümne binaarse sümmeetrilise tunnuse abil: $\boldsymbol{x_1} = (1,0,1,1,0,0,1,0,0,0)$, $\boldsymbol{x_2} = (0,1,0,0,1,0,0,0,0,0)$, $\boldsymbol{x_3} = (0,1,0,0,0,0,0,1,1,0)$, $\boldsymbol{x_4} = (1,1,0,0,1,0,1,1,0,0)$. Arvutades nende tunnuste Hammingu erinevust saadakse järgmised tulemused:

$$D(\boldsymbol{x_2}, \boldsymbol{x_3}) = 0{,}3, \qquad D(\boldsymbol{x_1}, \boldsymbol{x_4}) = 0{,}5.$$

Jaccardi erinevusmõõdu korral on aga tulemused teistsugused:

$$D(\boldsymbol{x_2}, \boldsymbol{x_3}) = 0{,}750, \qquad D(\boldsymbol{x_1}, \boldsymbol{x_4}) = 0{,}714.$$

Näeme, et Hammingu erinevuse kohaselt on vaatluste $\boldsymbol{x_1}$ ja $\boldsymbol{x_4}$ erinevus suurem kui vaatluste $\boldsymbol{x_2}$ ja $\boldsymbol{x_3}$ korral. Jaccardi erinevusmõõdu kohaselt on tulemus vastupidine, vaatluste $\boldsymbol{x_1}$ ja $\boldsymbol{x_4}$ erinevus on nüüd väiksem. Selline asjaolu võib viia erineva klasterduseni kahe erineva

erinevusmõõdu korral, seega on väga oluline kindlaks teha, kas tegemist on sümmeetriliste või asümmeetriliste tunnustega ning seejärel valida sobiv erinevusmõõt.

1.2 Sarnasusmõõdud enama kui kahe väärtusega kvalitatiivsete tunnuste korral

Olgu vaatlused $\boldsymbol{x_i}$ ja $\boldsymbol{x_j}$ kirjeldatud p tunnuse abil ja tunnuse l, $l = 1, \dots, p$, võimalikud väärtused on $1, \dots, m_l$. Kõige levinum viis antud vaatluste sarnasuse välja arvutamiseks on jälle kasutada lihtsat sarnasuskoefitsienti.

Definitsioon 3. *Kui kvalitatiivsetel tunnustel on rohkem kui kaks võimalikku väärtust, on lihtne sarnasuskoefitsient jälle defineeritud kui keskmine sarnasuste arv:*

$$S(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{1}{p}\sum_{l=1}^{p} S_{ijl},$$

kus

$$S_{ijl} = \begin{cases} 0, & kui\ \boldsymbol{x_i}\ ja\ \boldsymbol{x_j}\ väärtused\ on\ tunnuse\ l\ korral\ erinevad, \\ 1, & kui\ \boldsymbol{x_i}\ ja\ \boldsymbol{x_j}\ väärtused\ on\ tunnuse\ l\ korral\ samad. \end{cases}$$

Antud sarnasusmõõdule vastav erinevusmõõt näitab keskmist vaatlustevahelist erinevuste arvu.

Kui tegemist on järjestustunnusega, siis tähendaks lihtsa sarnasuskoefitsiendi kasutamine informatsiooni osalist kaotamist. Olgu etteantud samad vaatlused $\boldsymbol{x_i}$ ja $\boldsymbol{x_j}$, mis on kirjeldatud p järjestustunnuse abil. Iga tunnuse l, $l = 1, \dots, p$, korral on selle võimalikud väärtused $1, \dots, m_l$ järjestatud. Kui väärtused on järjestatud, siis mida lähemal on need üksteisele, seda sarnasemad nad on. Näiteks tunnuse „tervise seisund“ väärtusteks võivad olla „halb tervis“=1, „rahuldav tervis=2“, „hea tervis=3“ ja „suurepärane tervis=4“. Tervise seisundid „hea“ ja „suurepärane“ on üksteisele lähedal, seega need on ka sarnasemad, samal ajal seisundid „halb“ ja „hea“ on vägagi erinevad. Sellisel juhul tuleks arvestada ka väärtuste paaridega, mille korral on väärtused üksteisele lähedal.

Järjestustunnuste korral kasutatakse vaatlustevaheliste erinevuste mõõtmiseks samu kaugusmõõte nagu kvantitatiivsete tunnuste korral. Kaugusmõõtude kasutamiseks viiakse järjestustunnuse väärtused uuele skaalale nii, et l-nda tunnuse ja i-nda vaatluse esialgne väärtus r_{il}^* asendatakse uue väärtusega r_{il}:

$$r_{il} = \frac{r_{il}^* - 1}{m_l - 1}. \qquad (1)$$

Saadud uued väärtused on vahemikus [0,1] ja vaatlustevahelise kauguse mõõtmiseks saab kasutada näiteks eukleidilist või Manhattani kaugust.

Näide 2. Vaatleme 8 vaatlust, mis on kirjeldatud kolme järjestustunnuse põhjal: „tervise seisund“ väärtustega „halb tervis“=1, „rahuldav tervis“=2, „hea tervis“=3 ja „suurepärane tervis“=4; „sissetulek“ väärtustega „madal“=1, „keskmine“=2 ja „kõrge“=3; „haridustase“ väärtustega „põhiharidus“=1, „keskharidus“=2, „bakalaureus“=3 ja „magister“=4. Olgu vaatluste väärtused järgmised: $\boldsymbol{x_1} = (2,2,3)$, $\boldsymbol{x_2} = (1,3,4)$, $\boldsymbol{x_3} = (2,3,3)$, $\boldsymbol{x_4} = (3,1,1)$, $\boldsymbol{x_5} = (4,2,2)$, $\boldsymbol{x_6} = (3,1,1)$, $\boldsymbol{x_7} = (1,3,3)$, $\boldsymbol{x_8} = (1,2,4)$. Kuna tegemist on järjestustunnustega, siis viime saadud väärtused uuele skaalale kasutades valemit 1 (vt tabel 2). Näiteks vaatluse $\boldsymbol{x_1}$ uus väärtus tunnuse „tervise seisund“ korral on $\frac{2-1}{4-1} = 1/3$.

Tabel 2. Uuritavate vaatluste tunnuste väärtused uuel skaalal

	Tervis	Sissetulek	Haridus
$\boldsymbol{x_1}$	1/3	1/2	2/3
$\boldsymbol{x_2}$	0	1	1
$\boldsymbol{x_3}$	1/3	1	2/3
$\boldsymbol{x_4}$	2/3	0	0
$\boldsymbol{x_5}$	1	1/2	1/3
$\boldsymbol{x_6}$	2/3	0	0
$\boldsymbol{x_7}$	0	1	2/3
$\boldsymbol{x_8}$	0	1/2	1

Vaatlustevaheliste kauguste arvutamiseks kasutame eukleidilist kaugust. Saadud tulemuste põhjal koostame kauguste maatriksi, mida on näha joonisel 1.

```
     1    2    3    4    5    6    7
2 0.68
3 0.50 0.47
4 0.90 1.56 1.25
5 0.75 1.30 0.90 0.68
6 0.90 1.56 1.25 0.00 0.68
7 0.60 0.33 0.33 1.38 1.17 1.38
8 0.47 0.50 0.68 1.30 1.20 1.30 0.60
```

Joonis 1. Uuritavate vaatluste kauguste maatriks

Eeldame, et uuritavad objektid soovitakse jagada kahte klastrisse PAM-algoritmi abil. Selleks kasutame tarkvara R-i lisapaketi „*Cluster*“ funktsiooni *pam*, millest räägitakse peatükis 2.2. Maatriksist näeme, et kõige suurem objektidevaheline kaugus on vaatlustel $\boldsymbol{x_2}$ ja $\boldsymbol{x_4}$ ning $\boldsymbol{x_2}$ ja $\boldsymbol{x_6}$, mis võiks viidata sellele, et nende paaride objektid ei saa asuda ühes klastris. Paneme tähele, et vaatluste $\boldsymbol{x_4}$ ja $\boldsymbol{x_6}$ omavaheline kaugus on 0, st vaatluste väärtused on samad iga tunnuse korral, seega objektide paar peaks asuma samas klastris. Esimesse klastrisse sattusid vaatlused $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, $\boldsymbol{x_3}$, $\boldsymbol{x_7}$ ja $\boldsymbol{x_8}$ ning teise klastrisse vaatlused $\boldsymbol{x_4}$, $\boldsymbol{x_5}$ ja $\boldsymbol{x_6}$. Esimese klastri objektid on kõrgema hariduse, halvema tervise ja suurema sissetulekuga, samal ajal teise klastri objektid on madalama hariduse, parema tervise ja väiksema sissetulekuga.

2 K-medoidide meetod

Klasteranalüüsi eesmärk on leida klastrid nii, et sama klastri objektid oleksid võimalikult sarnased ja erinevate klastrite objektid võimalikult erinevad. Selleks võib kasutada näiteks erinevaid kaugusel põhinevaid klasterdusmeetodeid, mille hulgas leiab hierarhilisi ja tükeldamismeetodeid. Tükeldamismeetodite hulka kuuluvad K-medoidide ja K-keskmiste meetodid. Kui K-keskmiste meetodi korral on klastrit esindavaks objektiks kõigi objektide keskmine, siis K-medoidide meetodi korral otsitakse klastrit esindavat objekti ehk medoidi klastri objektide hulgast. Alapeatüki allikana kasutatakse raamatu Izenman (2008) peatükki 12 ja raamatu Kaufman ja Rousseeuw (1990) peatükki 2, kui ei ole märgitud teisiti.

2.1 K-medoidide meetodi ja PAM-algoritmi kirjeldus

K-medoidide meetodi korral leitakse klastrit esindav objekt ehk medoid ja paigutatakse ülejäänud vaatlused lähima medoidi juurde nii, et medoidi ja klastri objektide erinevuste summa oleks minimaalne. Seega sihifunktsioon ESS_{med}, mis sõltub eelnevalt määratud erinevusmõõdust, on defineeritud kui

$$ESS_{med} = \sum_{k=1}^{K} \sum_{c(i)=k} d_{ii_k},$$

kus $c(i)$ tähistab i-nda objekti klastrit ja $d_{ii_k} = d(\boldsymbol{x_i}, \boldsymbol{x_{i_k}})$ tähistab objektide $\boldsymbol{x_i}$ ja $\boldsymbol{x_{i_k}}$ omavahelist erinevust. Medoid $\boldsymbol{x_{i_k}}$ defineeritakse kui klastrisisene objekt, mis minimeerib sihifunktsiooni ESS_{med} väärtuse, st mille erinevus teiste klastri objektidega on minimaalseim:

$$i_k = arg\ min_{\{i:c(i)=k\}} \sum_{c(j)=k} d_{ij}.$$

PAM-algoritm (*partitioning around medoids)* ehk tükeldamine medoidide ümber on K-medoidide meetodi modifikatsioon. PAM-algoritmil ja K-medoidide meetodil on eesmärk ja sihifunktsioon samad, kuid algoritmid, mille abil jõutakse eesmärgini, on erinevad. Vaatleme antud meetodite algoritme lähemalt.

1. Määratud erinevusmõõdu kohaselt arvutame välja erinevuste maatriksi $D = (d_{ij})$.
2. Fikseerime klastrite arvu K ja moodustame esialgsed klastrid.
3. Iga klastri k jaoks, $k = 1, \ldots, K$, leiame medoidid.

4a. Vaatleme algoritmi jätku K-medoidide meetodi korral.

- Paigutame kõik objektid klastritesse vastavalt sellele, millisele medoidile on objekt kõige lähemal. Paneme tähele, et seejuures sihifunktsiooni ESS_{med} väärtus väheneb.
- Kordame sammu 3 ja 4a seni, kuni klasterdus jääb samaks.

4b. Vaatleme algoritmi jätku PAM-algoritmi korral.

- Iga medoidi ja iga vaatluse korral, mis ei ole medoid, kaalume, kas nende vahetamine toob kaasa sihifunktsiooni väärtuse vähenemise, st vajadusel paigutame medoidi ja teise vaatluse ümber, seejuures jälgides, et sihifunktsiooni ESS_{med} väärtuse vähenemine oleks maksimaalne.
- Kordame ümberpaigutamise protsessi nii kaua, kuni klasterdus jääb samaks.

PAM-algoritmis on kaks etappi: medoidide leidmise faas (nn *BUILD*-faas) ehk algoritmi osa, kus valitakse välja klastreid esindavad objektid, ja ümberpaigutamise faas (nn *SWAP*-faas), kus uuritakse, kas esmaselt valitud medoidide hulka on võimalik paremaks muuta sihifunktsiooni väärtuse vähendamise abil. Mõlemad PAM-algoritmi faasid mõjutavad lõplikku klasteranalüüsi tulemust, seega uurime neid lähemalt. Paneme tähele, et medoidide leidmise faas vastab ülaltoodud algoritmi sammudele 2 ja 3. Otsime võimalikult head medoidide komplekti.

1) Uurime medoidide leidmise faasi lähemalt.
    - Olgu välja valitud algne medoid, mille erinevus kõikidest teistest objektidest on minimaalne.
    - Iga vaatluse $\boldsymbol{x_i}$ korral, mis ei ole veel medoidiks välja valitud, ja iga suvalise objekti $\boldsymbol{x_j}$ korral arvutame nende objektide vahelise erinevuse $d(\boldsymbol{x_i}, \boldsymbol{x_j})$.

- Arvutame väärtuse $C_{ji} = max(D_j - d(\boldsymbol{x_i}, \boldsymbol{x_j}), 0)$, kus D_j on objekti $\boldsymbol{x_j}$ ja temale kõige lähema medoidi vaheline erinevus.
    a. Kui $C_{ji} = 0$, siis vaatluste $\boldsymbol{x_i}$ ja $\boldsymbol{x_j}$ vaheline erinevus on suurem kui D_j, st vaatluse $\boldsymbol{x_j}$ erinevus vaatlusest $\boldsymbol{x_i}$ on suurem temale lähimast medoidist ja vaatluse $\boldsymbol{x_i}$ valimine medoidi rolli ei ole soodne objekti $\boldsymbol{x_j}$ seisukohast.
    b. Kui $C_{ji} > 0$, siis vaatluste $\boldsymbol{x_i}$ ja $\boldsymbol{x_j}$ vaheline erinevus on väiksem kui D_j, st vaatlus $\boldsymbol{x_i}$ on vaatlusega $\boldsymbol{x_j}$ sarnasem kui vaatlus $\boldsymbol{x_j}$ ja temale lähim medoid, seega $\boldsymbol{x_i}$ valimine medoidi rolli on soodne objekti $\boldsymbol{x_j}$ seisukohast. Järelikult huvitutakse maksimaalsest C_{ji} väärtusest.
- Iga $\boldsymbol{x_i}$ korral (mis pole veel medoid) arvutame kõigi vaatluste $\boldsymbol{x_j}$ panuste summa $\sum_j C_{ji}$ ja valime välja uueks medoidiks sellise objekti $\boldsymbol{x_i}$, mille korral $\sum_j C_{ji}$ on maksimaalne, seega $\boldsymbol{x_i}$ on kõige soodsam medoidi kandidaat.
- Kordame ülaltoodud protsessi seni, kuni kõik K medoidi on leitud.

2) Ümberpaigutamise faasis vaatleme objektide paari $(\boldsymbol{x_i}, \boldsymbol{x_h})$, kus $\boldsymbol{x_i}$ on esimeses etapis välja valitud medoid ja $\boldsymbol{x_h}$ on uus medoidi kandidaat. Olgu $\boldsymbol{x_j}$ suvaline vaatlus, mis pole medoid. Arvutame väärtuse C_{jih}, mis näitab kui palju objekt $\boldsymbol{x_j}$ panustab vaatluste $\boldsymbol{x_i}$ ja $\boldsymbol{x_h}$ ümberpaigutamisse. Vaatluste $\boldsymbol{x_i}$ ja $\boldsymbol{x_j}$ paiknemiseks on kaks võimalust.
    - Kui vaatlused $\boldsymbol{x_i}$ ja $\boldsymbol{x_j}$ on ühes klastris, siis $d(\boldsymbol{x_j}, \boldsymbol{x_i}) = D_j$, st objektile $\boldsymbol{x_j}$ kõige lähimaks medoidiks ongi $\boldsymbol{x_i}$. Vaatleme erinevaid võimalusi objektide $\boldsymbol{x_j}$ ja $\boldsymbol{x_h}$ paiknemiseks üksteise suhtes.
        a. Olgu E_j erinevus vaatluse $\boldsymbol{x_j}$ ja temale teise lähima medoidi vahel. Kui vaatluse $\boldsymbol{x_j}$ erinevus temale teisest lähimast medoidist on suurem kui vaatlusest $\boldsymbol{x_h}$, st $d(\boldsymbol{x_j}, \boldsymbol{x_h}) < E_j$, siis $C_{jih} = d(\boldsymbol{x_j}, \boldsymbol{x_h}) - d(\boldsymbol{x_j}, \boldsymbol{x_i})$. Paneme tähele, et väärtus C_{jih} võib olla nii positiivne kui ka negatiivne.

Kui vaatluse $\boldsymbol{x_j}$ erinevus vaatlusest $\boldsymbol{x_h}$ on suurem kui vaatlusest $\boldsymbol{x_i}$, siis $C_{jih} > 0$ ja vaatluste $\boldsymbol{x_i}$ ja $\boldsymbol{x_h}$ vahetamine objekti $\boldsymbol{x_j}$ seisukohast ei ole soodne. Seega, kui $C_{jih} < 0$, siis vaatluse $\boldsymbol{x_j}$ erinevus vaatlusest $\boldsymbol{x_i}$ on suurem kui vaatlusest $\boldsymbol{x_h}$ ja vaatluste $\boldsymbol{x_i}$ ja $\boldsymbol{x_h}$ vahetamine on soodne.

b. Kui vaatluse $\boldsymbol{x_j}$ erinevus vaatlusest $\boldsymbol{x_h}$ on suurem või võrdne temale teise lähima medoidi erinevusega, st $d(\boldsymbol{x_j}, \boldsymbol{x_h}) \geq E_j$, siis $C_{jih} = E_j - D_j$, kusjuures väärtus C_{jih} on sellisel juhul alati positiivne, sest objektide $\boldsymbol{x_i}$ ja $\boldsymbol{x_h}$ vahetamine ei ole soodne.

- Kui aga vaatlused $\boldsymbol{x_i}$ ja $\boldsymbol{x_j}$ on erinevates klastrites, siis $d(\boldsymbol{x_j}, \boldsymbol{x_i}) > D_j$.

  c. Kui vaatluse $\boldsymbol{x_j}$ erinevus vaatlusest $\boldsymbol{x_h}$ on suurem kui erinevus talle kõige lähema medoidiga, st $d(\boldsymbol{x_j}, \boldsymbol{x_h}) > D_j$, siis väärtus $C_{jih} = 0$ ja vaatluste $\boldsymbol{x_i}$ ja $\boldsymbol{x_h}$ ümbervahetus objekti $\boldsymbol{x_j}$ seisukohast ei ole soodne, sest $\boldsymbol{x_j}$ ei anna mingit informatsiooni objektide $\boldsymbol{x_i}$ ja $\boldsymbol{x_h}$ ümbervahetamiseks.

  d. Kui vaatluse $\boldsymbol{x_j}$ erinevus temale kõige lähemast medoidist on suurem kui vaatlusest x_h, st $d(\boldsymbol{x_j}, \boldsymbol{x_h}) < D_j$, siis $C_{jih} = d(\boldsymbol{x_j}, \boldsymbol{x_h}) - D_j$, kusjuures väärtus C_{jih} on alati negatiivne, st vaatluste $\boldsymbol{x_i}$ ja $\boldsymbol{x_h}$ ümbervahetus objekti $\boldsymbol{x_j}$ seisukohast on alati soodne.

- Iga paari $(\boldsymbol{x_i}, \boldsymbol{x_h})$ korral arvutame kõigi vaatluste $\boldsymbol{x_j}$ panuste summa $T_{ih} = \sum_j C_{jih}$ ja valime välja sellise paari $(\boldsymbol{x_i}, \boldsymbol{x_h})$, mille korral on T_{ih} minimaalne.
- Kui väärtus T_{ih} on negatiivne, mis garanteerib sihifunktsiooni ESS_{med} väärtuse vähenemise, siis vahetame objektid $\boldsymbol{x_i}$ ja $\boldsymbol{x_h}$ omavahel ära ja lähme ümberpaigutamise faasi algusesse. Kui aga T_{ih} on positiivne või võrdub nulliga, siis algoritm peatub, st objektide $\boldsymbol{x_i}$ ja $\boldsymbol{x_h}$ ümbervahetus ei too kaasa sihifunktsiooni vähenemist.

Näide 3. Vaatleme näidet 2, kus on tehtud klasterdus järjestustunnuste „tervise seisund“, „sissetulek“ ja „haridustase“ korral, ja uurime saadud tulemust põhjalikumalt. Järgmiseks vaatleme PAM-algoritmi rakendamist kasutades tarkvara *R* funktsiooni *pam* ja selle argumenti $trace.lev$, mille abil saab illustreerida PAM-algoritmi ümberpaigutamise faasi (vt peatükk 2.2).

Valime algseteks medoidideks suvaliselt vaatlused $\boldsymbol{x_2}$ ja $\boldsymbol{x_5}$ ning vaatleme vaatlustevahelisi kaugusi nende medoididega. Medoid, mille korral on vaatlustevaheline kaugus minimaalne, on vaatlusele lähim medoid. Tabelis 3 on näha, et lähim medoid vaatlustele $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \boldsymbol{x_7}$ ja $\boldsymbol{x_8}$ on $\boldsymbol{x_2}$ ja vaatlustele $\boldsymbol{x_4}, \boldsymbol{x_5}$ ja $\boldsymbol{x_6}$ on medoid $\boldsymbol{x_5}$. Sihifunktsiooni väärtus ESS_{med} medoidide $\boldsymbol{x_2}$ ja $\boldsymbol{x_5}$ korral on 3,34.

Tabel 3. Vaatlustevahelised kaugused väljavalitud medoididega

		$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_4}$	$\boldsymbol{x_5}$	$\boldsymbol{x_6}$	$\boldsymbol{x_7}$	$\boldsymbol{x_8}$
Samm 1	Erinevused medoidist x_2	0,68	0	0,47	1,56	1,3	1,56	0,33	0,5
	Erinevused medoidist x_5	0,75	1,3	0,9	0,68	0	0,68	1,17	1,2
	Lähim medoid	$\boldsymbol{x_2}$	$\boldsymbol{x_2}$	$\boldsymbol{x_2}$	$\boldsymbol{x_5}$	$\boldsymbol{x_5}$	$\boldsymbol{x_5}$	$\boldsymbol{x_2}$	$\boldsymbol{x_2}$
Samm 2	Erinevused medoidist x_2	0,68	0	0,47	1,56	1,3	1,56	0,33	0,5
	Erinevused medoidist x_4	0,9	1,56	1,25	0	0,68	0	1,38	1,3
	Lähim medoid	$\boldsymbol{x_2}$	$\boldsymbol{x_2}$	$\boldsymbol{x_2}$	$\boldsymbol{x_4}$	$\boldsymbol{x_4}$	$\boldsymbol{x_4}$	$\boldsymbol{x_2}$	$\boldsymbol{x_2}$
Samm 3	Erinevused medoidist x_4	0,9	1,56	1,25	0	0,68	0	1,38	1,3
	Erinevused medoidist x_7	0,6	0,33	0,33	1,38	1,17	1,38	0	0,6
	Lähim medoid	$\boldsymbol{x_7}$	$\boldsymbol{x_7}$	$\boldsymbol{x_7}$	$\boldsymbol{x_4}$	$\boldsymbol{x_4}$	$\boldsymbol{x_4}$	$\boldsymbol{x_7}$	$\boldsymbol{x_7}$

Järgmise sammuna vahetakse vana medoid $\boldsymbol{x_5}$ uue medoidi $\boldsymbol{x_4}$ vastu, sest et uue medoidi ja medoidi $\boldsymbol{x_2}$ korral on sihifunktsiooni ESS_{med} väärtus 2,66 ehk võrreldes eelmise sammuga väiksem. Tabelis 3 näeme, et medoidi $\boldsymbol{x_4}$ ja vaatluse $\boldsymbol{x_6}$ omavaheline kaugus on 0, mis vähendab sihifunktsiooni väärtust 0,68 võrra (sest medoidi $\boldsymbol{x_2}$ korral väärtused jäävad samaks).

Ümberpaigutamise faasi viimase sammuna vahetatakse ära vana medoid $\boldsymbol{x_2}$ uue medoidi $\boldsymbol{x_7}$ vastu. Saadud medoidide komplekti ESS_{med} väärtus on 2,54 ja see summa enam väiksemaks ei lähe (vt joonis 1).

Näeme, et klasterduse tulemusena on ühes klastris vaatlused $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \boldsymbol{x_7}$ ja $\boldsymbol{x_8}$, (klastrit esindavaks objektiks on $\boldsymbol{x_7}$) ja teises klastris vaatlused $\boldsymbol{x_4}, \boldsymbol{x_5}$ ja $\boldsymbol{x_6}$ (esindavaks objektiks on

$\boldsymbol{x_4}$). Esimese klastri „esindavaks objektiks“ on halva tervise, kõrge sissetuleku ja bakalaureuse haridusega isik ning teist klastrit esindab hea tervise, madala sissetuleku ja põhiharidusega isik. Paneme tähele, et kuna vaatlused $\boldsymbol{x_4}$ ja $\boldsymbol{x_6}$ on võrdsed, siis medoidide paari $\boldsymbol{x_6}$ ja $\boldsymbol{x_7}$ sihifunktsiooni ESS_{med} väärtus on samuti 2,54, seega käesolev vaatluste paar sobib samuti lõplikuks medoidide komplektiks.

Kuigi K-medoidide meetodi algoritm on sarnane K-keskmiste meetodi algoritmiga, üks K-medoidide meetodi eelistest on meetodi üldisus. K-medoidide meetodi korral on võimalus kasutada üldiseid erinevusmõõte, samal ajal K-keskmiste meetod on defineeritud eukleidilise kauguse jaoks. Samuti K-medoidide meetodi eeliseks on stabiilsus ehk robustsus, sest et meetodi sihifunktsioon ei ole defineeritud läbi eukleidilise kauguse ruudu, mis on väga tundlik erindite suhtes. Seega võib eeldada, et K-medoidide meetod toimib erindite korral hästi ja saadud tulemus on paremini interpreteeritav. Klastrid, mis on saadud K-medoidide meetodi tulemusena, võivad olla nii sfäärilised kui ka mittesfäärilised, kuid K-keskmiste meetodi korral on üldjuhul võimalik saada ainult sfäärilisi klastreid.

K-medoidide meetod ja PAM-algoritm töötavad suurepäraselt väikeste andmestike korral, kuid nende meetodite rakendamine võtab palju aega, kui andmestik on suurem. Sellisel juhul saab rakendada CLARA-algoritmi (*Clustering Large Applications*), mille eesmärk on täpselt sama nagu PAM-algoritmil. CLARA-algoritm kasutab analüüsimiseks ainult osa andmetest. Täpsemalt, moodustatakse osavalim kõikidest klasterdavatest objektidest juhuslike arvude generaatori abil ja teostatakse klasteranalüüs PAM-algoritmi abil. Seejärel kasutatakse osavalimi medoide ja paigutatakse kogu andmestiku objektid lähima osavalimi medoidi juurde, st toimub kõikide objektide klasterdamine. Protsessi korratakse mitu korda ja valitakse välja selline klasterdus, mille korral on sihifunktsiooni väärtus minimaalne. Selline algoritm võimaldab säästa programmi arvutamise aega ja kasutada tarkvara mälu väiksemas mahus.

Näide 4. Järgmine näide illustreerib K-medoidide meetodi rakendamist kvalitatiivsete tunnuste korral. Näide on ära toodud raamatu Hennig jt (2016) neljandas peatükis. Antud analüüsi eesmärk oli klasterdada 100 looma ja lindu (mille hulgas leidub ka inimtüdruk) järgmiste binaarsete tunnuste alusel (väärtus 1 tähistab tunnuse olemasolu): karvad, suled, munad, piim, lendamisoskus, side veega, kiskja, hambad, selgroog, hingamisoskus, mürk, saba, uimed,

kabjad ja sarv. Samuti mõõdeti objektide jalgade arvu. Klasteranalüüs on läbi viidud K-medoidide meetodi abil. Vaatlused klasterdati 13 klastrisse, mis on välja toodud koos klastri medoidiga järgmises loetelus:

1. metsikud imetajad (ja inimtüdruk): gepard, hunt, ilves, inimtüdruk, karu, kass, **leopard**, lõvi, mangust, metssiga, mutt, naarits, nokkloom, opossum, pesukaru, puuma, tuhkur, tuhnik;
2. kodustatud imetajad: **hamster**, kits, lehm, merisiga, poni, põhjapõder;
3. röövlinnud: kiivi, **kull**, nandu, raisakotkas, vares;
4. kalad 1: ahven, astelrai, **haug**, heeringas, koerkala, merimadu, piraaja, säga, teib, tuunikala;
5. putukad: kilpkonn, **kirp**, lepatriinu, nälkjas, sipelgas, sääsk, uss;
6. kalad 2: karpkala, **kiltursk**, kärnkonn, merihobu, merikeel;
7. merelinnud: **kajakas**, mustviires, pingviin, änn;
8. mitteröövloomad: antiloop, elevant, gorilla, **hirv**, jänes, kaelkirjak, känguru, nahkhiir, orav, orüks, piison, suur-vereimeja, uruhiir;
9. lendavad putukad: herilane, mesilane, **toakärbes**, ööliblikas;
10. selgrootud veeloomad: homaar, **jõevähk**, merekarp, kaheksajalg, krabi, meduus, meritäht;
11. roomajad: konn, rästik, salamander, skorpion, **tuataara**, vaskuss;
12. linnud: faasan, flamingo, jaanalind, kana, käblik, luik, lõoke, papagoi, part, tuvi, **varblane**;
13. veeimetajad: **delfiin**, hüljes, merilõvi, pringel.

Paneme tähele, et tekkinud klastrid on hästi interpreteeritavad. Üldiselt on iga klastri objektid sarnased ja kuuluvad ühte loomade või lindude klassi (nt merelinnud, kodustatud imetajad jne), kusjuures inimtüdruk on paigutatud metsikute imetajate hulka, mis on selle objekti jaoks kõige

sobilikum valik. Siiski näeme, et kalade klastreid on kaks ja ei ole võimalik täpselt öelda, mis tunnuste põhjal on mõlema klastri kalad eraldatud. Märkame, et kilpkonn on paigutatud putukate klastrisse, kuigi ta kuulub roomajate klassi. Lepatriinu on lendav putukas, aga antud klasteranalüüsi tulemusena oli ta paigutatud teise putukate klassi. Nii kärnkonn kui ka uss peaksid asuma roomajate klastris, kuid klasterdamise tulemusena nad on paigutatud klastritesse kalad 2 ja putukad.

Samad objektid olid klasterdatud ka modifitseeritud K-keskmiste meetodiga (*OCKM* ehk *order-constrained K-means clustering*), tulemust näeb artiklis Steinley ja Hubert (2008). Kuigi kvalitatiivsete tunnuste korral K-keskmiste meetodi rakendamine ei ole üldjuhul õige, on antud juhul kõik tunnused binaarsed (välja arvatud jalgade arv) ja eukleidilise kauguse ruutude summa koosneks ainult ühtedest ja nullidest, st eukleidilise kauguse ja K-medoidide meetodi puhul kasutatavad erinevusmõõdud annavad klasterdamisel sama tulemuse. Võrreldes kahe meetodi korral saadud tulemusi võib tähele panna, et K-keskmiste meetodi korral on tekkinud kabjaliste klaster, kuhu kuulub enamik K-medoidide meetodi kodustatud imetajate klastri objektidest. Samuti on tekkinud kahepaikseliste klaster, mille objektid on K-medoidide meetodi korral laiali paigutatud. Märgime, et modifitseeritud K-keskmiste meetodi tulemusena on tekkinud ainult üks kalade klaster, samaaegselt K-medoidide meetodi korral on kalad jaotatud kahte klastrisse.

Kokkuvõtteks sõltub klasterduse tulemus oluliselt meetodi valikust ja klasterduse eesmärgist.

2.2 PAM-algoritm tarkvaras R

Selleks, et rakendada PAM-algoritmi tarkvaras R, kasutatakse lisapaketi „*Cluster*“ funktsiooni *pam*. Kasutusele on võetud lisapaketi versioon 2.0.6.

Funktsioon *pam* klasterdab vaatlused PAM-algoritmi alusel K klastrisse. Selle funktsiooni üks tähtsamatest argumentidest on x, mille abil saab ette anda andmestiku või erinevuste maatriksi. Kui tegemist on erinevuste maatriksiga, siis tuleb seda funktsioonis näidata, määrates argumendi $diss$ väärtuseks $TRUE$. Klastrite arvu määramiseks on argument k ja erinevusmõõdu määramiseks argument $metric$. Paneme tähele, et erinevuste maatriksi kasutamise korral argumenti $metric$ täpsustamine ei ole vajalik. Kui tegu on tavalise

andmemaatriksiga, siis saab kasutada eukleidilist ja Manhattani kaugust. Juhul, kui soovitakse medoidide hulka ette anda, kasutatakse argumenti $medoids$. Sellisel juhul medoidide leidmise faasi ei toimu. Kui tahetakse arvutada objektide omavahelisi erinevusi standardiseeritud andmete põhjal, siis tuleb seda funktsioonis näidata, määrates argumendi $stand$ väärtuseks $TRUE$. Kuna antud töös tegeletakse kvalitatiivsete andmetega, siis vaatluste standardiseerimine ei ole vajalik.

Selleks, et uurida detailsemalt PAM-algoritmi iteratsioone, on võimalik kasutada funktsiooni *pam* argumenti $trace.lev$, mille väärtuseks sobib positiivne täisarv, mis määrab soovitud iteratsioonide arvu. Argument väljastab algsete medoidide komplekti, näitab ümberpaigutamise faasi samme ja nendega kaasnevaid minimaalseid kaugusi väljavalitud medoidide korral. Samuti väljastatakse sihifunktsiooni väärtused iga medoidide komplekti korral.

3 Mudelipõhine klasteranalüüs

Järgnevalt kirjeldame mudelipõhise klasteranalüüsi teooriat kvalitatiivsete tunnuste korral. Mudelipõhise klasteranalüüsi korral eeldatakse, et klasterdatavad andmed on saadud mitmemõõtmelise segujaotuse abil, kusjuures kvalitatiivsete tunnuste korral on selleks multinomiaalsete jaotuste segu. Allikatena on kasutatud raamatut Hennig jt (2016) ja magistritööd Mirski (2019), kui ei ole viidatud teisiti.

3.1 Mudelipõhise klasteranalüüsi kirjeldus

Olgu klasterdatavad vaatlused $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ kirjeldatud p kvalitatiivse tunnuse abil. Igal tunnusel l on m_l võimalikku väärtust. Soovime paigutada vaatlused $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ paarikaupa lõikumatutesse gruppidesse ehk klastritesse $C_1, C_2, \ldots, C_K$. Mudelipõhise klasteranalüüsi korral kasutatakse vaatluste jaotuse kirjeldamiseks mitmemõõtmelist segujaotust.

Definitsioon 4. *Olgu Z latentne juhuslik suurus võimalike väärtustega $1, \ldots, K$, mille tõenäosused on $\mathbb{P}\{Z = k\} = \pi_k$, $k = 1, \ldots, K$. Öeldakse, et p-mõõtmeline juhuslik vektor $\boldsymbol{X}$ on juhuslike komponentide $\boldsymbol{X}_1, \ldots, \boldsymbol{X_K}$ segu, kui selle tihedusfunktsioon avaldub kujul*

$$f(\boldsymbol{x};\, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \, f_k(x;\, \boldsymbol{\theta}_k), \tag{2}$$

kus $\boldsymbol{x} \in \mathbb{R}^p$, $\pi_1, \ldots, \pi_K$ on komponentide kaalud, $\pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$, f_k on komponendi $\boldsymbol{X}_k$ tihedusfunktsioon, $\boldsymbol{\theta}_k$ on selle tiheduse parameetrite vektor ning $\boldsymbol{\theta} = \{\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ tähistab segujaotuse (2) kõigi parameetrite hulka.

Ülaltoodud definitsioonis mõiste „latentne suurus“ tähendab suurust, mida ei saa otse mõõta ja mida hinnatakse teiste seotud tunnuste abil. Näiteks „tervis“ on latentne suurus, mille hinnangut saab määrata peale järgmisi mõõtmisi: kehakaal, vererõhk, veresuhkur, keha temperatuur jne.

Kui tegeletakse kvantitatiivsete tunnustega, siis võetakse kasutusele üldjuhul mitmemõõtmelise normaaljaotusega komponentide segu. Vaatluste klasterdamiseks kvalitatiivsete tunnuste korral sobib aga K-komponendiline mitmemõõtmeliste multinomiaalsete jaotuste segu. Selleks, et rakendada multinomiaalset segujaotust eeldame, et vaatluse $\boldsymbol{x_i} = (x_i^1, \ldots, x_i^p)'$ l-nda tunnuse väärtust kirjeldatakse binaarse vektori $(x_i^{l1}, x_i^{l2}, \ldots, x_i^{lm_l})'$ abil, kus x_i^{lh}=1, kui sellel tunnusel on h-s võimalik väärtus, ja $x_i^{lh} = 0$

vastasel juhul, $l = 1,\dots,p$, $i = 1,\dots,n$. Seega iga vaatlus on avaldatav binaarse vektori $\boldsymbol{x} = (x^{11},\dots,x^{1m_1};\ \dots;\ x^{p1},\dots,x^{pm_p})'$ kaudu, mille tõenäosusfunktsioon on järgmine:

$$f(\boldsymbol{x};\ \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k\, \mathcal{M}_k(\boldsymbol{x};\ \boldsymbol{\alpha}_k) = \sum_{k=1}^{K} \pi_k \prod_{l=1}^{p}\prod_{h=1}^{m_l} (\alpha_k^{lh})^{x^{lh}}, \qquad (3)$$

kus $\pi_1,\dots,\pi_K$ on komponentide kaalud, $\pi_k \geq 0$, $\sum_{k=1}^{K}\pi_k = 1$, $\mathcal{M}_k(\boldsymbol{x};\ \boldsymbol{\alpha}_k)$ on mitmemõõtmelise multinomiaalse jaotuse tõenäosusfunktsioon, α_k^{lh} on tõenäosus, et l-ndal tunnusel on h-s võimalik väärtus, kui x on selle segujaotuse k-nda komponendi realisatsioon, $\boldsymbol{\alpha}_k = (\alpha_k^{11},\dots,\alpha_k^{1m_1};\ \dots;\ \alpha_k^{p1},\dots,\alpha_k^{pm_p})'$, $\sum_{h=1}^{m_l}\alpha_k^{lh} = 1,\ k = 1,\dots,K,\ l = 1,\dots,p,$ ja $\boldsymbol{\theta} = \{\pi_1,\dots,\pi_K,\boldsymbol{\alpha_1},\dots,\boldsymbol{\alpha_K}\}$.

Mitmemõõtmeliste multinomiaalsete jaotuste segu korral eeldatakse, et kvalitatiivsed tunnused on lokaalselt sõltumatud, st nad on sõltumatud iga komponendi sees. Lokaalse sõltumatuse peamiseks põhjuseks on lihtsa sõltuvusnäitaja puudumine kvalitatiivsete tunnuste korral. Kui kvantitatiivsete tunnuste korral kasutatakse kovariatsiooni tunnuste sõltuvuse mõõtmiseks, siis sama lihtsat ja kergesti tõlgendatavat sõltuvusnäitajat kvalitatiivsete tunnuste jaoks pole olemas. Kuigi võib arvata, et sõltumatus on väga kitsendav eeldus, klasteranalüüsi korral töötab see üldjuhul hästi ja annab häid tulemusi. Sõltumatuse eelduse tõttu saab avaldise (3) multinomiaalse jaotuse tõenäosusfunktsiooni $\mathcal{M}_k(\boldsymbol{x};\ \boldsymbol{\alpha}_k)$ esitada korrutisena üle tunnuste klastri sees.

3.2 Segujaotuse parameetrite hindamine

Multinomiaalsete jaotuste segu (3) parameetrite hindamiseks kasutatakse EM-algoritmi (*expectation-maximization algorithm*), mis on suurima tõepära meetodi iteratiivne modifikatsioon latentsete tunnuste korral. Defineeritakse hulk $\boldsymbol{z} = \{\boldsymbol{z_1},\dots,\boldsymbol{z_n}\}$, mis koosneb juhuslikest indikaatorvektoritest $\boldsymbol{z_i} = (z_{i1},\dots,z_{iK})'$. Kui $z_{ik} = 1$, siis vaatlus $\boldsymbol{x_i}$ on genereeritud segujaotuse komponendi $\boldsymbol{X_k}$ abil, vastasel juhul aga $z_{ik} = 0$, $i = 1,\dots,n$, $k = 1,\dots,K$. Olgu $\boldsymbol{Z} = \{\boldsymbol{Z_1},\dots,\boldsymbol{Z_n}\}$ elemendid $\boldsymbol{Z_i}$ vektoritele $\boldsymbol{z_i}$ vastavad multinomiaalse jaotusega juhuslikud vektorid. Soovime maksimeerida logaritmilist tõepärafunktsiooni

$$ln\ p(\boldsymbol{x};\boldsymbol{z};\boldsymbol{\theta}) = ln\left[\prod_{i=1}^{n}\prod_{k=1}^{K}[\pi_k f_k(\boldsymbol{x_i};\boldsymbol{\theta_k})]^{z_{ik}}\right],$$

kuid see ei ole võimalik $\boldsymbol{z}$ tundmatuse tõttu. Tõepärafunktsiooni

$$ln\, p(\boldsymbol{x};\boldsymbol{\theta}) = ln\left[\prod_{i=1}^{n} f(\boldsymbol{x_i};\boldsymbol{\theta})\right] = \sum_{i=1}^{n} ln\left(\sum_{k=1}^{K} \pi_k \prod_{l=1}^{p}\prod_{h=1}^{m_l} (\alpha_k^{lh})^{x^{lh}}\right)$$

maksimeerimine ei ole samuti võimalik, sest leitud parameetrite hinnangud sisaldavad kaudselt vektoreid $\boldsymbol{z_i}$. Seega arvutatakse tinglikku keskväärtust $E[ln\, p(\boldsymbol{x};\boldsymbol{Z};\boldsymbol{\theta})|\boldsymbol{x}]$ vektori $\boldsymbol{Z_i}$ tingliku jaotuse suhtes.

Algoritmi teostatakse kahes etapis: E-samm ja M-samm. Enne E-sammu fikseeritakse segujaotuse „vanad parameetrid" ehk algväärtused $\boldsymbol{\theta^{vana}}$, täpsemalt π_k^{vana} ja $\boldsymbol{\alpha}_k^{vana}$, $k = 1,\ldots,K$. E-sammu ehk keskväärtustamise sammu käigus leitakse $E[ln\, p(\boldsymbol{x};\boldsymbol{Z};\boldsymbol{\theta})|\boldsymbol{x};\ \boldsymbol{\theta^{vana}}]$, mis antud olukorras taandub tinglikute tõenäosuste $\gamma_k(\boldsymbol{x_i};\ \boldsymbol{\theta^{vana}})$ arvutamiseks. Tinglik tõenäosus $\gamma_k(\boldsymbol{x_i};\boldsymbol{\theta})$ näitab, kui tõenäoliselt on vaatlus $\boldsymbol{x_i}$ multinomiaalsete jaotuste segu k-nda komponendi realisatsioon:

$$\gamma_k(\boldsymbol{x_i};\boldsymbol{\theta}) = \mathbb{P}\{Z_{ik} = 1|\boldsymbol{X} = \boldsymbol{x_i}\} = \frac{\pi_k f_k(\boldsymbol{x_i};\boldsymbol{\theta_k})}{f(\boldsymbol{x_i};\boldsymbol{\theta})}.$$

M-sammu ehk maksimeerimise sammu käigus leitakse multinomiaalsete jaotuste segu parameetrite hinnangud $\boldsymbol{\theta^{uus}}$ ehk parameetrite väärtused, mis maksimeerivad keskväärtustamisel saadud funktsiooni. Hinnangud $\boldsymbol{\theta^{uus}}$ ei pruugi olla lõplikud, sest tuleb kontrollida EM-algoritmi koondumist. Kui algoritm ei koondu, siis protsessi korratakse.

Kõige levinum on logaritmilise tõepära suhtelise muutuse koondumiskriteerium

$$\frac{ln\, p(\boldsymbol{x};\ \boldsymbol{\theta^{uus}}) - ln\, p(\boldsymbol{x};\ \boldsymbol{\theta^{vana}})}{|ln\, p(\boldsymbol{x};\ \boldsymbol{\theta^{vana}})|} < \varepsilon, \qquad (4)$$

kus $\varepsilon > 0$ on mingi väike väärtus. Algoritm peatub, kui logaritmilise tõepära suhteline muutus on väiksem kui ε või kui on saavutatud etteantud iteratsioonide arv. Tarkvaras R kasutatakse mudelipõhise klasterdamise teostamisel sama koondumiskriteeriumit. Kui algoritm ei koondu, siis minnakse uuesti E-sammu juurde, võttes saadud parameetrite hinnangud uuteks algväärtusteks, $\boldsymbol{\theta^{uus}} \rightarrow \boldsymbol{\theta^{vana}}$**.** Üldjuhul tuleb rakendada EM-algoritmi mitu korda heade hinnangute leidmiseks kasutades erinevaid alglähendeid.

Kvalitatiivsete tunnuste korral on EM-algoritmi tulemusena saadavad multinomiaalsete jaotuste segu parameetrite hinnangud järgmised:

$$\pi_k^{uus} = \frac{n_k}{n}, \quad \alpha_k^{lh\ uus} = \frac{1}{n_k}\sum_{i=1}^{n} \gamma_k(\boldsymbol{x_i},\boldsymbol{\theta^{vana}}) x_i^{lh},$$

kus $n_k = \sum_{i=1}^{n} \gamma_k(\boldsymbol{x_i},\boldsymbol{\theta^{vana}})$, $k = 1,\ldots,K$, $l = 1,\ldots,p$, $h = 1,\ldots,m_l$. Komponentide kaalude hinnangute π_k^{uus} korral vaadeldakse ligikaudset vaatluste arvu klastris ja kogu

vaatluste arvu suhet. Tõenäosuse hinnangu $\alpha_k^{lh\ uus}$ korral summeeritakse tinglikud tõenäosused üle vaatluste kokku tunnuse l h-nda väärtuse korral ja jagatakse ligikaudse vaatluste arvuga klastris.

3.3 Hinnatavate segumudelite klassid

Defineerime uuritavad parameetrid $\boldsymbol{\alpha_k^l} = (\alpha_k^{l1}, \ldots, \alpha_k^{lm_l})'$ kasutades moodi tõenäosust ehk suurimat tõenäosust igas klastris iga tunnuse korral. Olgu moodi tõenäosuseks väärtus γ_k^l, siis on tõenäosuste vektor $\boldsymbol{\alpha_k^l}$ kujul $(\beta_k^l, \ldots., \beta_k^l, \gamma_k^l, \beta_k^l, \ldots, \beta_k^l)'$, kus $\beta_k^l = (1 - \gamma_k^l)/(m_l - 1)$ ja $\beta_k^l < \gamma_k^l, k = 1, \ldots, K, l = 1, \ldots, p$. Selline parametriseerimine, kus iga klastri ja iga tunnuse korral on üks tõenäosus teistest suurem ja ülejäänud tõenäosused on võrdsed, annab võimaluse seada multinomiaalsete jaotuste segule kitsendusi, mille abil on võimalik hinnatavate parameetrite arvu vähendada.

Olgu $h(k, l) \in \{1, \ldots, m_l\}$ tõenäosuse γ_k^l positsioon tõenäosuste vektoris $\boldsymbol{\alpha_k^l}$. Et kirjeldada vektori $\boldsymbol{\alpha_k^l}$ väärtusi kasutame parameetreid a_k^{lh}, mille korral $a_k^{lh} = 1$, kui $h = h(k, l)$, ja $a_k^{lh} = 0$ vastasel juhul. Seega saame vektori $\boldsymbol{a_k^l} = (a_k^{l1}, \ldots, a_k^{lm_l})'$ ning multinomiaalsete jaotuste uued parameetrid kitsenduste korral avalduvad kujul

$$\alpha_k^{lh} = \begin{cases} 1 - \varepsilon_k^l, & kui\ h = h(k, l), \\ \varepsilon_k^l/(m_l - 1), & vastasel\ juhul, \end{cases}$$

kus $\varepsilon_k^l = 1 - \gamma_k^l$.

Ülaltoodud parameetriseerimist kasutades saab vaadelda viit segumudelite klassi:

- standardne kitsendusteta segumudel $[\varepsilon_k^{lh}]$, mille parameetrid sõltuvad nii klastrist, tunnusest kui ka tunnuse võimalikest väärtustest;
- kitsendustega segumudel $[\varepsilon_k^l]$, mille moodi tõenäosus $\gamma_k^l = 1 - \varepsilon_k^l$ sõltub nii klastrist kui ka tunnusest;
- kitsendustega segumudel $[\varepsilon_k]$, mille moodi tõenäosus γ_k^l sõltub ainult klastrist, $\gamma_k^l = \gamma_k$;
- kitsendustega segumudel $[\varepsilon^l]$, mille moodi tõenäosus γ_k^l sõltub ainult tunnusest, $\gamma_k^l = \gamma^l$;
- kitsendustega segumudel $[\varepsilon]$, mille moodi tõenäosus γ_k^l ei sõltu ei klastrist ega tunnusest.

Juhul, kui tegemist on binaarsete tunnustega, $m_l = 2$, siis segumudel $[\varepsilon_k^{lh}]$ taandub mudeliks $[\varepsilon_k^{l}]$.

Paneme tähele, et standardse segumudeli $[\varepsilon_k^{lh}]$ korral tuleb hinnata $(K-1) + K\sum_l (m_l - 1)$ parameetrit, kuid kitsendustega mudeli $[\varepsilon_k^{l}]$ hinnatavate parameetrite arv muutub oluliselt väiksemaks, $(K-1) + Kp$. Näiteks, kui $K = 3$ ja vaadeldavad vaatlused on kirjeldatud viie tunnuse abil, $p = 5,\ m_1 = \ldots = m_5 = 4$, siis mudeli $[\varepsilon_k^{lh}]$ parameetrite arv on 47, samal ajal kitsendustega mudelil $[\varepsilon_k^{l}]$ on 17 parameetrit. Märgime, et lisaks tuleb arvestada ka moodi tõenäosuse positsiooni hindamisega. Kui vaatlused on kirjeldatud ainult binaarsete tunnuste abil, siis on mõlema mudeli parameetrite arv sama.

3.4 Integreeritud klassifitseerimistõepära kriteerium

Selleks, et välja valida optimaalseim klasterdus ehk parim mudel mudelipõhise klasteranalüüsi korral, võib rakendada integreeritud klassifitseerimistõepära kriteeriumit, mis põhineb Bayesi informatsioonikriteeriumil (BIC). Bayesi kriteerium sõltub mudeli maksimiseeritud tõepärast ning karistusliikmest, mis sõltub nii mudeli parameetrite arvust kui ka valimimahust.

Definitsioon 5. *Bayesi informatsioonikriteerium põhineb mudeli maksimiseeritud tõepära logaritmil:*

$$BIC = -2ln\, L(\widehat{\boldsymbol{\theta}}) + \nu\, ln\, n,$$

kus $L(\widehat{\boldsymbol{\theta}})$ on mudeli maksimiseeritud tõepära, ν on parameetrite arv mudelis ja n on valimimaht.

Käesolevat kriteeriumit on võimalik samuti rakendada parima mudeli tuvastamiseks, kuid on märgatud, et Bayesi kriteerium ei ole väga sobilik just klasteranalüüsi jaoks. Integreeritud klassifitseerimistõepära kriteerium ICL_{bic} aga võtab arvesse asjaolu, et uuritavad segumudelid on hinnatud mudelipõhise klasteranalüüsi teostamise eesmärgil. Paneme tähele, et nii BIC kui ka ICL_{bic} korral on parimaks mudeliks see, mille kriteeriumi väärtus on minimaalne.

Definitsioon 6. *Integreeritud klassifitseerimistõepära kriteerium ehk ICL_{bic} on defineeritud kui*

$$ICL_{bic} = BIC - 2\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{\gamma}_k(\boldsymbol{x_i};\boldsymbol{\theta})\, ln\, \hat{\gamma}_k(\boldsymbol{x_i};\boldsymbol{\theta}),$$

kus $\hat{\gamma}_k(\boldsymbol{x_i};\boldsymbol{\theta}) = \gamma_k(\boldsymbol{x_i};\widehat{\boldsymbol{\theta}})$ *on hinnatud tinglik tõenäosus, et vaatlus* $\boldsymbol{x_i}$ *on multinomiaalsete jaotuste segu* k*-nda komponendi realisatsioon.*

Kvalitatiivsete tunnuste korral kasutatakse ka nn täpset ICL kriteeriumit. Võttes nii komponentide kaalude $\pi_1, \dots, \pi_K$ kui ka parameetrite $\boldsymbol{\alpha_k^l}$ eelmõõduks mitteinformatiivse Dirichlet jaotuse, saab nn integreeritud täistõepära $p(\boldsymbol{x};\boldsymbol{z})$, mis tuleneb $p(\boldsymbol{x};\boldsymbol{z};\boldsymbol{\theta})$ integreerimisest üle $\boldsymbol{\theta}$ jaotuse, multinomiaalsete jaotuste segu korral täpselt välja arvutada. Seega on nn täpne ICL kriteerium defineeritud järgmiselt.

Definitsioon 7. *Täpne ICL kriteerium avaldub kujul*

$$ICL = ln\, p(\boldsymbol{x},\hat{\boldsymbol{z}}) = \sum_{k=1}^{K} ln\, \Gamma\left(\hat{n}_k + \frac{1}{2}\right) + \sum_{k=1}^{K}\sum_{l=1}^{p}\left[\sum_{h=1}^{m_l} ln\, \Gamma\left(\hat{u}_k^{lh} + \frac{1}{2}\right) - ln\, \Gamma\left(\hat{n}_k + \frac{m_l}{2}\right)\right]$$

$$+ln\, \Gamma\left(\frac{K}{2}\right) - K\, ln\, \Gamma\left(\frac{1}{2}\right) - ln\, \Gamma\left(n\ + \frac{K}{2}\right) + K\sum_{l=1}^{p}\left[ln\, \Gamma\left(\frac{m_l}{2}\right) - m_l\, ln\, \Gamma\left(\frac{1}{2}\right)\right],$$

kus $\hat{n}_k = \sum_{i=1}^{n} \hat{z}_{ik}$, $\hat{u}_k^{jh} = \sum_{i=1}^{n} \hat{z}_{ik} x_i^{jh}$, Γ *tähistab gammafunktsiooni ja*

$$\hat{z}_{ik} = \begin{cases} 1, & kui\ k = arg\ max_h\, \hat{\gamma}_h(\boldsymbol{x_i};\boldsymbol{\theta}) \\ 0, & vastasel\ juhul \end{cases}.$$

Täpse ICL kriteeriumi kohaselt on parimaks mudeliks see, mille kriteeriumi väärtus on maksimaalne. Paneme tähele, et tarkvaras R mudelipõhise klasteranalüüsi teostamiseks kasutatavas funktsioonis arvutatakse ainult ICL_{bic} väärtust.

4 Kriteeriumid klasterduste võrdlemiseks

4.1 Randi indeks

Järgnevalt antakse ülevaade *Randi indeksist* ja tema kohandatud versioonist, mille abil saab võrrelda kahte erinevat klasterdust, mis on teostatud sama andmestiku põhjal. Antud indeksi korral ei ole oluline, kuidas ja mis meetodiga on mingi klasterdus saadud. Samuti ei eeldata, et saadud klastrite arv oleks võrdne. Peatüki allikana kasutatakse artiklit Hubert ja Arabie (1985).

Randi indeks on kriteerium, mis mõõdab, kui hästi on erinevate meetodite abil saadud klasterdused kooskõlas. Vaatleme kahte klasterdust n vaatluse jaoks: $C = \{C_1, \dots, C_K\}$ ja $C' = \{C'_1, \dots, C'_K,\}$. Vaatleme kõiki võimalike vaatluste paare $(\boldsymbol{x_i}, \boldsymbol{x_j})$ ning objektide $\boldsymbol{x_i}$ ja $\boldsymbol{x_j}$ paiknemist klasterdustes C ja C':

a) olgu a selliste paaride arv, mille korral paari vaatlused kuuluvad samasse klastrisse nii klasterduse C kui ka C' korral;

b) olgu b selliste paaride arv, mille korral paari vaatlused kuuluvad erinevatesse klastritesse nii klasterduse C kui ka C' korral;

c) olgu c selliste paaride arv, mille korral paari vaatlused kuuluvad erinevatesse klastritesse klasterduse C korral, aga samasse klastrisse klasterduse C' korral;

d) olgu d selliste paaride arv, mille korral paari vaatlused kuuluvad samasse klastrisse klasterduse C korral, aga erinevatesse klastritesse klasterduse C' korral.

Seega uurime, kas erinevad paarid $(\boldsymbol{x_i}, \boldsymbol{x_j})$ on kooskõlas kahe erineva klasterduse puhul. Juhtude a) ja b) korral on vaatluste paarid kooskõlas. Paneme tähele, et kokku on $\binom{n}{2}$ paari.

Definitsioon 8. *Randi indeks on defineeritud kui tõenäosus, et suvaline objektide paar on kooskõlas klasterduste C ja C' korral:*

$$RI = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}.$$

Mida suurem on kooskõlas paaride arv ja väiksem vastuolus paaride arv, seda enam on kaks uuritavat klasterdust kooskõlas Randi indeksi järgi. Randi indeksi puuduseks on asjaolu, et indeks ei arvesta juhuslikkuse komponendiga. Kuna Randi indeks ei võta arvesse klasterduse

meetodit ja klastrite arvu, siis on raske indeksi suurust tõlgendada ja võrrelda. Näiteks, kui ühe klasterduse korral on palju klastreid ja klasterdavaid objekte ning teise klasterduse korral on nii klastreid kui ka objekte vähe, siis on raske saadud Randi indeksi väärtusi omavahel võrrelda. Võttes arvesse indeksi keskväärtust nullolukorra puhul (indeksi keskmine väärtus olukorras, kus vaatlused on paigutatud klastritesse juhuslikult) saame normaliseeritud indeksi ehk *kohandatud Randi indeksi* (ARI ehk *Adjusted Rand Index*).

Vaatleme klasterduste C ja C' võrdlemiseks järgnevat sagedustabelit (vt tabel 4). Väärtus n_{ij}, $i = 1, \dots, K$, $j = 1, \dots, K'$, näitab ühiste vaatluste arvu klastrite C_i ja C'_j korral. Tähistagu väärtus $n_{i\cdot}$, $i = 1, \dots, K$, klastri C_i kõikide objektide arvu ja väärtus $n_{\cdot j}$, $j = 1, \dots, K'$, klastri C'_j kõikide objektide arvu.

Tabel 4. Sagedustabel klasterduste C ja C' võrdlemiseks

Klaster	C'_1	C'_2	…	$C'_{K'}$	Summa
C_1	n_{11}	n_{12}	…	$n_{1K'}$	$n_{1\cdot}$
C_2	n_{21}	n_{22}	…	$n_{2K'}$	$n_{2\cdot}$
⋮	⋮	⋮	⋱	⋮	⋮
C_K	n_{K1}	n_{K2}	…	$n_{KK'}$	$n_{K\cdot}$
Summa	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot K'}$	n

Kohandatud Randi indeksi korral lahutatakse Randi indeksist selle keskväärtus ja jagatakse indeksi maksimumi ja keskväärtuse erinevusega ehk jagatakse vahemiku pikkusega, kus paiknevad huvipakkuvad indeksi väärtused:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathrm{E(RI)}}{1 - \mathrm{E(RI)}}.$$

Kuna Randi indeksi keskväärtus kirjeldab selle indeksi keskmist väärtust olukorras, kus vaatlused oleksid paigutatud klastritesse $C_1, .., C_K$ ja $C'_1, \dots, C'_{K'}$ juhuslikult, ei ole me huvitatud väärtustest, mis on väiksemad kui selle indeksi keskväärtus.

Definitsioon 9. *Kohandatud Randi indeks mõõdab erinevust Randi indeksi keskväärtusest olukorras, kus klasterdused on saadud juhuslikult:*

$$ARI = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K'}\binom{n_{ij}}{2} - [\sum_{i=1}^{K}\binom{n_{i\cdot}}{2}][\sum_{j=1}^{K'}\binom{n_{\cdot j}}{2}]/\binom{n}{2}}{[\sum_{i=1}^{K}\binom{n_{i\cdot}}{2} + \sum_{j=1}^{K'}\binom{n_{\cdot j}}{2}]/2 - [\sum_{i=1}^{K}\binom{n_{i\cdot}}{2}][\sum_{j=1}^{K'}\binom{n_{\cdot j}}{2}]/\binom{n}{2}} =$$

$$= \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}.$$

Kohandatud Randi indeks on 0, kui Randi indeks võrdub selle keskväärtusega, maksimum saavutatakse väärtuses 1. Huvipakkuvad ARI väärtused on vahemikus [0,1], kuid kohandatud Randi indeks võib omada ka negatiivseid väärtusi, mis ei oma sisulist mõtet.

4.2 Keskmise silueti laiuse kriteerium

Tähistagu $C = \{C_1, \ldots, C_K\}$ objektide $\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}$ uuritavat klasterdust ja olgu $d(\boldsymbol{x_i}, \boldsymbol{x_j})$ objektide $\boldsymbol{x_i}$ ja $\boldsymbol{x_j}$ omavaheline erinevus. Olgu vaatlus $\boldsymbol{x_i}$ paigutatud klastrisse C_k. *Keskmise silueti laiuse kriteerium* (*average silhouette width*) sobib optimaalsete klastrite arvu leidmiseks erinevusmõõtudel põhinevate klasterdusmeetodite korral ning näitab, kui soodne on vaatluste klasterdamine üldiselt (Kaufman ja Rousseeuw, 1990).

Definitsioon 10. *Keskmise silueti laiuse kriteerium on väärtus*

$$\bar{s}_K(C) = \frac{1}{n}\sum_{i=1}^{n} s_i = \frac{1}{n}\sum_{i=1}^{n}\frac{b_i - a_i}{max\{a_i, b_i\}},$$

kus

$$a_i = \frac{1}{|C_k| - 1}\sum_{j \in C_k, i \neq j} d(\boldsymbol{x_i}, \boldsymbol{x_j})$$

ja

$$b_i = min_{l \neq k}\frac{1}{|C_l|}\sum_{j \in C_l} d(\boldsymbol{x_i}, \boldsymbol{x_j}).$$

Silueti laius s_i näitab, kui hästi on vaatlus $\boldsymbol{x_i}$ paigutatud, $s_i \in [-1, 1]$. Väärtus a_i tähistab vaatluse $\boldsymbol{x_i}$ keskmist erinevust oma klastri C_k ülejäänud vaatlustest. Seega a_i näitab, kui hästi vaatlus $\boldsymbol{x_i}$ sobib klastrisse C_k. Mida väiksem on a_i, seda sarnasem on vaatlus $\boldsymbol{x_i}$ klastri C_k teiste

vaatlustega. Väärtus b_i tähistab aga vaatluse $\boldsymbol{x_i}$ keskmist erinevust temale järgmise lähima klastri vaatlustest. Klaster, mille keskmine erinevus objektiga $\boldsymbol{x_i}$ on minimaalne, on vaatluse $\boldsymbol{x_i}$ naaberklaster. Kui $a_i \ll b_i$, siis objekt $\boldsymbol{x_i}$ on oma klastri C_k objektidele keskmiselt lähemal kui naaberklastri vaatlustele ja s_i väärtus on 1 lähedal. Kui aga $b_i \ll a_i$, siis ei sobi vaatlus $\boldsymbol{x_i}$ oma klastrisse C_k ja s_i väärtus on -1 lähedal. Kui a_i ja b_i on enam vähem võrdsed, siis s_i väärtus on 0 lähedal, st ei saa kindlalt öelda kas vaatluse $\boldsymbol{x_i}$ jaoks sobib paremini oma klaster C_k või naaberklaster. Seega mida suurem on $\bar{s}_K$ väärtus, seda optimaalsem on klastrite arv K.

5 Simulatsioonide näited

Mudelipõhise klasteranalüüsi ja PAM-algoritmi võrdlemiseks viiakse läbi simulatsioonid erinevate olukordade korral. Alguses uuritakse meetodite käitumist ühe andmestiku korral, seejärel aga teostatakse klasteranalüüs saja andmestiku korral.

Näide 5. Käesoleva näite eesmärk on võrrelda PAM-algoritmi ja mudelipõhist klasteranalüüsi kvalitatiivsete tunnuste korral. Heaks klasteranalüüsi meetodiks kattuvate klastrite korral peetakse mudelipõhist klasteranalüüsi, kuid ei ole teada, millist võrreldavatest meetoditest oleks parem kasutada mingi konkreetse olukorra puhul. Genereerime vaatlusi kolmekomponendilisest segujaotusest (3), $K = 3$, vaatluste arv on $n = 500$. Vaatlused on kirjeldatud nelja tunnuse abil, $p = 4$, millel on vastavalt $m_1 = m_2 = 2$, $m_3 = m_4 = 3$ võimalikku väärtust. Et genereerida erineva klastrite kattuvuse määraga andmestikke, rakendame parameetrite määramiseks järgmist valemit (Biernacki, Celeux ja Govaert, 2010, lk 2996):

$$\alpha_k^{lh} = \begin{cases} \frac{1}{m_l} + (1-\delta)\frac{m_l - 1}{m_l}, & kui\ h = [(k-1)\ mod\ m_l] + 1, \\ \frac{1 - \frac{1}{m_l} - (1-\delta)\frac{m_l - 1}{m_l}}{m_l - 1}, & vastasel\ juhul, \end{cases}$$

kus $\delta \in [0,1]$ näitab kattuvuse määra ja funktsioon mod tähistab jäägiga jagamist, $l = 1, \dots p$, $h = 1, \dots, m_j$, $k = 1, \dots, K$. Kui $\delta = 0$, on kattuvus minimaalne ja saame järgmised parameetrite vektorid kolmekomponendilise segujaotuse korral:

$$\alpha_1 = (1;0;\ \ 1;0;\ \ 1;0;0;\ \ 1;0;0)',$$
$$\alpha_2 = (0;1;\ \ 0;1;\ \ 0;1;0;\ \ 0;1;0)',$$
$$\alpha_3 = (1;0;\ \ 1;0;\ \ 0;0;1;\ \ 0;0;1)'.$$

Juhul, kui $\delta = 1$, on kattuvus maksimaalne ja $\alpha_k^{lh} = {}^1/_{m_l}$, st iga tunnuse kõik väärtused on sama tõenäosusega iga klastri korral:

$$\alpha_1 = \alpha_2 = \alpha_3 = \left(\frac{1}{2};\frac{1}{2};\ \ \frac{1}{2};\frac{1}{2};\ \ \frac{1}{3};\frac{1}{3};\frac{1}{3};\ \ \frac{1}{3};\frac{1}{3};\frac{1}{3}\right)'.$$

Selliste parameetritega saadud segumudel vastab kitsendustega mudelile $[\varepsilon^l]$, sest et sagedasema väärtuse tõenäosus γ_k^l on igas klastris võrdne, $\gamma_k^l = \gamma^l$, st moodi tõenäosus ei sõltu klastrist. Seega klasteranalüüsi käigus vaadeldakse ainult kitsendustega mudeleid $[\varepsilon_k^l]$, $[\varepsilon^l]$ ja $[\varepsilon]$.

Vaatluste genereerimisel kasutame nii erinevaid kui ka võrdseid komponentide kaalusid. Uurime olukordi, kui $\pi_1 = 0{,}15, \pi_2 = 0{,}35$ ja $\pi_3 = 0{,}5$ ning $\pi_1 = \pi_2 = \pi_3 = {}^1/_3$. Samuti vaatleme erinevaid klastrite kattuvuse määrasid $\delta = 0{,}2$ ja $\delta = 0{,}6$. Kokkuvõttes saame neli erinevat andmestikku, millega teostame klasteranalüüsi klastrite arvu $K = 2, 3, 4$ jaoks.

Klasteranalüüsi teostamiseks PAM-algoritmi abil kasutatakse tarkvara *R* lisapaketi *„Cluster“* funktsiooni *pam.* Vaatlustevaheliste erinevuste mõõtmiseks PAM-algoritmi korral kasutatakse lihtsal sarnasuskoefitsiendil põhinevat erinevusmõõtu, milleks on keskmine erinevuste arv. PAM-algoritmi korral valitakse parim klasterdus keskmise silueti laiuse kriteeriumi $\bar{s}_K$ põhjal. Mudelipõhise klasteranalüüsi jaoks kasutatakse lisapaketi *„Rmixmod“* funktsiooni *mixmodCluster*. Mudeli parameetrite hindamiseks kasutatakse EM-algoritmi korrates protsessi 40 korda, kus maksimaalseks iteratsioonide arvuks on 1000 ja koondumiskriteeriumi (4) konstant on $\varepsilon = 0{,}0001$. Et valida välja parim mudel mudelipõhise klasterdamise korral, rakendatakse täpset ICL kriteeriumit. Parim mudel on see, mille korral on täpne ICL kriteerium maksimaalne. Antud näites mõõdetakse $\bar{s}_K$ ka mudelipõhise klasterdamise korral meetodite võrdlemise jaoks. Selleks, et võrrelda uuritavate meetodite tulemusi, võetakse appi ka mudelipõhise klasteranalüüsi ja PAM-algoritmi vahelist ARI kriteeriumit.

Enne, kui hakkame uurima tulemusi kattuvuste $\delta = 0{,}2$ ja $\delta = 0{,}6$ korral, vaatleme mis juhtub kui $\delta = 0$. Tulemused on ära toodud tabelis 5. Ühelt poolt, kui vaadelda kolmekomponendiliste ja neljakomponendiliste segumudelite ICL väärtusi, siis näeme, et need on samad, st hinnatakse kolmekomponendiline mudel. Tõepoolest, kui vaatame neljakomponendilise segumudeli kaalude hinnanguid, siis näeme, et üks hinnang on 0. Keskmise silueti laius on võrdne ühega kolme klastriga segumudelite korral, mis samuti viitab mõlema klasterduse optimaalsusele.

Tabel 5. Klasteranalüüsi tulemused minimaalse klastrite kattuvuse korral, $\delta = 0$

k	Kaalud	$\bar{s}_K$ mudel	$\bar{s}_K$ PAM	ARI	ICL	ICL_{bic}
2	erinevad	0,893	0,892	1	-698,2	1566,4
3	erinevad	1	1	1	-550,5	1016,4
4	erinevad	1	0,494	0,996	-550,5	1115,8
2	võrdsed	0,833	0,833	1	-821,2	2078,7
3	võrdsed	1	1	1	-609,1	1131,1
4	võrdsed	1	0,672	0,997	-609,1	1230,6

Teiselt poolt, vaadeldes $\mathrm{ICL_{bic}}$ väärtusi ei saa öelda, et hinnatav mudel kolme ja nelja klastri korral on tegelikult sama. Kuna antud olukorras klastrid ei kattu, siis peaksid mõlemad klasterduse meetodid töötama sama hästi ja tulemused seda näitavadki. Paneme tähele, et parima segumudeli valimiseks tuleks vaadata ka saadud parameetrite hinnanguid.

Tulemused kattuvuse $\delta = 0{,}2$ korral. Klasteranalüüsi tulemused kattuvuse $\delta = 0{,}2$ korral on ära toodud tabelis 6. Tegelikud segumudeli parameetrid antud kattuvuse korral on järgmised:

$$\alpha_1 = (0{,}9;\ 0{,}1;\ \ 0{,}9;\ 0{,}1;\ \ 0{,}867;\ 0{,}067;\ 0{,}067;\ \ 0{,}867;\ 0{,}067;\ 0{,}067)',$$
$$\alpha_2 = (0{,}1;\ 0{,}9;\ \ 0{,}1;\ 0{,}9;\ \ 0{,}067;\ 0{,}867;\ 0{,}067;\ \ 0{,}067;\ 0{,}867;\ 0{,}067)',$$
$$\alpha_3 = (0{,}9;\ 0{,}1;\ \ 0{,}9;\ 0{,}1;\ \ 0{,}067;\ 0{,}067;\ 0{,}867;\ \ 0{,}067;\ 0{,}067;\ 0{,}867)'.$$

Uurides saadud klasteranalüüsi tulemusi väikese kattuvuse korral $\delta = 0{,}2$ saame, et nii võrdsete kui ka erinevate kaalude puhul on ICL järgi parim segumudel kolme klastriga. Erinevate kaalude korral on parimaks mudeliks kitsendustega segumudel $[\varepsilon^l]$, mille korral moodi tõenäosus sõltub ainult tunnusest, samade kaalude korral aga $[\varepsilon]$, mille korral moodi tõenäosus ei sõltu ei klastrist ega tunnusest. Parima segumudeli $[\varepsilon]$ parameetrite hinnangud on väga lähedased parameetrite tegelikele väärtustele:

$$\hat{\alpha}_1 = (0{,}864;\ 0{,}137;\ \ 0{,}864;\ 0{,}137;\ \ 0{,}864;\ 0{,}068;\ 0{,}068;\ \ 0{,}864;\ 0{,}068;\ 0{,}068)',$$
$$\hat{\alpha}_2 = (0{,}137;\ 0{,}864;\ \ 0{,}137;\ 0{,}864;\ \ 0{,}068;\ 0{,}864;\ 0{,}068;\ \ 0{,}068;\ 0{,}864;\ 0{,}068)',$$
$$\hat{\alpha}_3 = (0{,}864;\ 0{,}137;\ \ 0{,}864;\ 0{,}137;\ \ 0{,}068;\ 0{,}068;\ 0{,}864;\ \ 0{,}068;\ 0{,}068;\ 0{,}864)'.$$

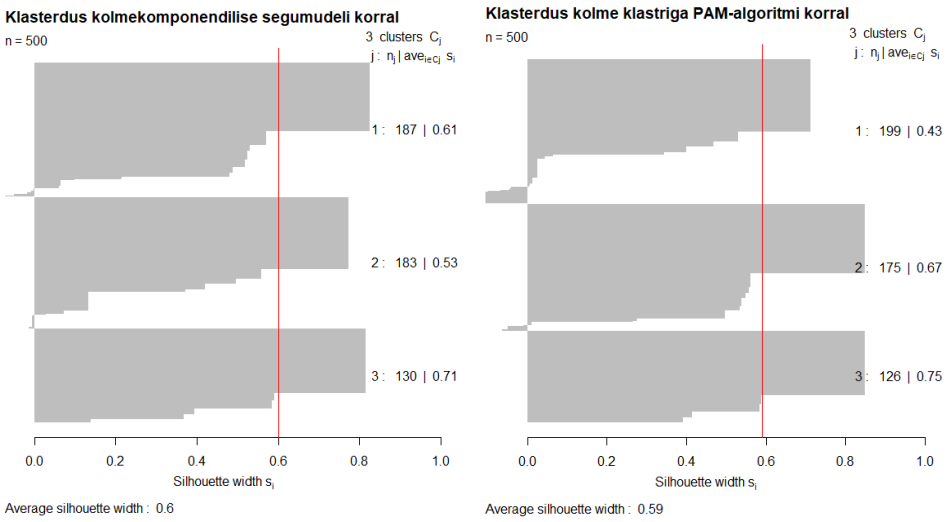
Võrdsete kaaludega andmestiku korral saadud parima mudeli kaalude hinnangud on $\hat{\pi}_1 = 0{,}329$, $\hat{\pi}_2 = 0{,}353$ ja $\hat{\pi}_3 = 0{,}318$ ja klasterduse veamäär on 10%.

Tabel 6. Klasteranalüüsi tulemused väikese klastrite kattuvuse korral, $\delta = 0{,}2$

k	Kaalud	$\bar{s}_K$ mudel	$\bar{s}_K$ PAM	ARI	ICL	ICL_{bic}
2	erinevad	0,643	0,633	0,853	-1369,5	2927,5
3	erinevad	0,614	0,610	0,681	-1350,4	2751,4
4	erinevad	0,616	0,604	0,643	-1351,2	2789,7
2	võrdsed	0,594	0,590	0,921	-1460,4	3251,5
3	võrdsed	0,605	0,592	0,878	-1406,8	2892,0
4	võrdsed	0,591	0,584	0,759	-1416,9	2950,8

Paneme tähele, et komponentide samade kaalude korral ka keskmise silueti laiuse kriteeriumi kohaselt on kolmekomponendiline mudel kõige soodsam, $\bar{s}_K = 0{,}605$. Erinevate kaalude korral aga on optimaalseim kahekomponendiline segumudel, $\bar{s}_K = 0{,}643$.

PAM-algoritmi korral on keskmise silueti laius kõikide situatsioonide korral väga sarnane, kuid erinevate kaalude korral on kriteeriumi väärtus kõige suurem kahekomponendilise klasterduse puhul, $\bar{s}_K = 0{,}633$, ja võrdsete kaalude korral kolmekomponendilise klasterduse puhul, $\bar{s}_K = 0{,}592$.



Joonis 2. Klasterdused kolme komponendiga mudelipõhise klasteranalüüsi ja PAM-algoritmi korral, $\delta = 0{,}2$

Joonisel 2 on esitatud kolmekomponendiliste klasterduste keskmise silueti laiuse graafikud võrdsete komponentide kaalude ja $\delta = 0{,}2$ korral. Joonisel näeme, et kõik klastrid on umbes sama laiusega ning ei leidu ühtegi klastrit, mille kõikide objektide silueti laiused oleksid alla keskmise silueti laiuse, st mõlemad klasterdused on sobilikud keskmise silueti laiuse kriteeriumi kohaselt. Samuti võib märgata, et PAM-algoritmi korral on ülemises klastris kõige rohkem objekte, mille silueti laius on negatiivne ehk mis sobiksid paremini naaberklastrisse.

Kui võrrelda klasterdamise tulemusi PAM-algoritmi ja mudelipõhise analüüsi korral, siis kõige rohkem on kooskõlas klasterdused kahe komponendiga nii erinevate kui ka samade kaalude korral ARI kohaselt. Paneme tähele, et ülal vaadeldud kolmekomponendiliste klasterduste puhul on ARI samuti kõrge. Märgime, et juba väikese klastrite kattuvuse korral meetodite tulemused eristuvad.

Tulemused kattuvuse $\delta = 0{,}6$ korral. Klasteranalüüsi tulemused kattuvuse $\delta = 0{,}6$ korral on ära toodud tabelis 7. Tegelikud segumudeli parameetrid antud kattuvuse korral on

$$\alpha_1 = (0{,}7;\ 0{,}3;\ \ 0{,}7;\ 0{,}3;\ \ 0{,}6;\ 0{,}2;\ 0{,}2;\ \ 0{,}6;\ 0{,}2;\ 0{,}2)',$$
$$\alpha_2 = (0{,}3;\ 0{,}7;\ \ 0{,}3;\ 0{,}7;\ \ 0{,}2;\ 0{,}6;\ 0{,}2;\ \ 0{,}2;\ 0{,}6;\ 0{,}2)',$$
$$\alpha_3 = (0{,}7;\ 0{,}3;\ \ 0{,}7;\ 0{,}3;\ \ 0{,}2;\ 0{,}2;\ 0{,}6;\ \ 0{,}2;\ 0{,}2;\ 0{,}6)'.$$

Vaadeldes saadud tulemusi suurema klastrite kattuvuse korral $\delta = 0{,}6$ saame, et nii erinevate kui ka samade kaalude korral on parimateks mudeliteks ICL järgi kahe klastriga segumudelid. Erinevate komponentide kaalude korral on parim kitsendustega mudel $[\varepsilon_k^l]$, mille korral sõltub moodi tõenäosus nii tunnusest kui ka klastrist, ning võrdsete kaalude korral mudel $[\varepsilon^l]$, mille korral sõltub moodi tõenäosus ainult tunnusest. Vaatleme saadud hinnanguid kolmekomponendilise segumudeli $[\varepsilon^l]$ korral:

$$\hat{\alpha}_1 = (0{,}729;\ 0{,}271;\ \ 0{,}716;\ 0{,}284;\ \ 0{,}556;\ 0{,}222;\ 0{,}222;\ \ 0{,}616;\ 0{,}192;\ 0{,}192)',$$
$$\hat{\alpha}_2 = (0{,}271;\ 0{,}729;\ \ 0{,}284;\ 0{,}716;\ \ 0{,}222;\ 0{,}556;\ 0{,}222;\ \ 0{,}192;\ 0{,}616;\ 0{,}192)',$$
$$\hat{\alpha}_3 = (0{,}729;\ 0{,}271;\ \ 0{,}716;\ 0{,}284;\ \ 0{,}222;\ 0{,}222;\ 0{,}556;\ \ 0{,}192;\ 0{,}192;\ 0{,}616)'.$$

Antud mudeli korral on kaalude hinnangud $\hat{\pi}_1 = 0{,}356$, $\hat{\pi}_2 = 0{,}254$ ja $\hat{\pi}_3 = 0{,}391$ ning klasterduse veamäär on 35,6%.

Tabel 7. Klasteranalüüsi tulemused suurema klastrite kattuvuse korral, $\delta = 0{,}6$

k	Kaalud	$\bar{s}_K$ mudel	$\bar{s}_K$ PAM	ARI	ICL	ICL_{bic}
2	erinevad	0,287	0,336	0,188	-1810,9	3809,9
3	erinevad	0,231	0,276	0,216	-1853,3	4108,1
4	erinevad	0,315	0,279	0,381	-1865,0	4361,8
2	võrdsed	0,217	0,297	0,280	-1787,0	3718,2
3	võrdsed	0,290	0,253	0,325	-1882,1	4222,5
4	võrdsed	0,203	0,277	0,391	-1857,4	4340,1

Kui vaadelda $\bar{s}_K$ mudelipõhise klasteranalüüsi võrdsete kaalude korral, siis vaatamata suuremale klastrite kattuvuse määrale on optimaalseimaks variandiks kolme klastriga segumudel, $\bar{s}_K = 0{,}290$. Erinevate kaalude korral on kõige soodsam aga neljakomponendiline

segumudel, $\bar{s}_K = 0{,}315$. Üldjuhul võib tähele panna, et mida suurem on klastrite arv mudelipõhise klasteranalüüsi korral, seda rohkem väiksemaid klastreid tekib. Näiteks antud kattuvuse korral samade kaaludega neljakomponendilise segumudeli saadud klastrites on 150, 180, 147 ja 23 objekti, kus viimane klaster on märgatavalt väiksem kui teised. PAM-algoritmi korral on optimaalseim klasterdus samuti kahe klastriga nii erinevate kui ka võrdsete kaalude korral, $\bar{s}_K = 0{,}336$ ja $\bar{s}_K = 0{,}297$ vastavalt. Kõige rohkem on kooskõlas neljakomponendilised klasterdused, $\mathrm{ARI} = 0{,}381$ erinevate kaalude korral ja $\mathrm{ARI} = 0{,}391$ võrdsete kaalude korral.

Näeme, et mida väiksem on klastrite arv, seda väiksem on ARI väärtus. Vaatleme erinevate kaalude korral saadud kahekomponendilisi klasterdusi, mille korral on mudelipõhisel analüüsil $\bar{s}_K = 0{,}287$ ja PAM-algoritmil $\bar{s}_K = 0{,}336$. Mudelipõhise klasterduse korral on objektide arv klastrites 115 ja 385, kuid PAM-algoritmi puhul on 256 ja 244. Seega antud olukorras tekitab PAM-algoritm võrdsema objektide arvuga klastreid, mis võibki põhjustada väikesi ARI väärtusi. Samuti võib tähele panna, et erinevate kaaludega kahekomponendiliste klasterduste korral mudelipõhise klasteranalüüsi puhul pannakse kokku esimene ja kolmas tegelik komponent, kuid PAM-algoritmi puhul on kokku pandud esimene ja teine komponent.

Kui võrrelda tulemusi kattuvuste $\delta = 0{,}2$ ja $\delta = 0{,}6$ korral, siis kahe meetodi klasterduste kooskõla ARI järgi on väikese kattuvuse korral suurem, st meetodid teostavad klasteranalüüsi suurema klastrite kattuvuse korral erinevalt.

Näide 6. Kuna ühe genereeritud andmestiku korral on raske mingeid põhjalikumaid järeldusi teha, siis kordame klasteranalüüsi saja erineva andmestiku puhul. Artiklis Anderlucci ja Hennig (2014) viiakse samuti klasteranalüüs läbi mudelipõhise analüüsi ja PAM-algoritmi abil kasutades väiksema (100, 200 ja 500) ja suurema (1000) vaatluste arvuga andmestikke väiksema ja suurema klastrite eralduse korral. Et analüüsi tulemused oleksid põhjalikumad ja täpsemad, vaadeldakse 4 ja 12 kvalitatiivset tunnust, mille võimalike väärtuste arv on 2, 4 või 8, ja komponentide arvuks valitakse 2, 3 ja 5. Antud magistritöös vaadeldakse aga klasteranalüüsi tulemusi ainult saja erineva andmestiku korral näite 5 stsenaariumi põhjal. Lisaks kattuvusele $\delta = 0{,}2$ ja $\delta = 0{,}6$ uuritakse kattuvusega $\delta = 0{,}4$ andmestike nii erinevate kui ka võrdsete kaalude korral. Klasteranalüüsi teostatakse klastrite arvu $K = 3$ jaoks ja

kasutatakse funktsiooni *pam* PAM-algoritmi jaoks ja funktsiooni *mixmodCluster* mudelipõhise klasterdamise jaoks. EM-algoritmi teostatakse 100 korda, iteratsioonide maksimaalseks arvuks on 1000 ja koondumiskriteeriumi (4) konstandiks on $\varepsilon = 0{,}0001$. Kuna tarkvara R lisapaketis „*Rmixmod*" kasutatakse parima segumudeli valimiseks $\mathrm{ICL}_{\mathrm{bic}}$ kriteeriumit, siis antud situatsioonis tegime otsuse selle kriteeriumi kasuks. Parim mudel on see, mille kriteeriumi väärtus on minimaalne. Et võrrelda saja simuleeritud andmestiku põhjal saadud klasteranalüüsi tulemusi, kasutatakse tegelikele klassidele vastavat keskmise silueti laiust $\bar{s}_K$, mudelipõhisele klasterdusele vastavat keskmise silueti laiust $\bar{s}_K$, PAM-algoritmi klasterdusele vastavat keskmise silueti laiust $\bar{s}_K$, võrreldakse mudelipõhise analüüsi ja PAM-algoritmi klasterdusi ARI abil, lisaks mõõdetakse ARI ka mudelipõhise analüüsi ja vaatluste tegelike klasside ning PAM-algoritmi ja vaatluste tegelike klasside võrdlemiseks.

Tulemused komponentide erinevate kaalude korral. Klasteranalüüsi tulemused erinevate kaalude korral, $\pi_1 = 0{,}15, \pi_2 = 0{,}35$ ja $\pi_3 = 0{,}5$, on ära toodud tabelis 8. Paneme tähele, et mida suurem on klastrite kattuvus, seda väiksemaks lähevad uuritavate kriteeriumite keskmised, st suurema kattuvuse korral on klastreid raskem eraldada. Kui võrrelda keskmise silueti laiuse kriteeriumite keskmisi, siis näeme, et suurema kattuvuse korral on nende väärtused PAM-algoritmi puhul pisut suuremad kui mudelipõhise klasteranalüüsi puhul. Lisaks sellele on ka standardhälbed PAM-algoritmi korral väiksemad, mis viitab meetodi stabiilsusele. Kui kattuvus on väiksem, $\delta = 0{,}2$, siis on nii mudelipõhise analüüsi kui ka PAM-algoritmi $\bar{s}_K$ umbes sama. Kuna aga keskmise silueti laiuse kriteerium põhineb erinevustel ja seda kasutatakse just PAM-algoritmi korral, siis on saadud tulemused oodatud. Märgime, et tegelikele klassidele vastav $\bar{s}_K$ keskmine on teistest keskmise silueti laiuse kriteeriumite keskmistest väiksem iga δ puhul.

Tabel 8. Klasteranalüüsi tulemuste keskmised ja standardhälbed 100 andmestiku ja erinevate kaalude korral

δ	$\bar{s}_K$ tegelik	$\bar{s}_K$ mudel	$\bar{s}_K$ PAM	ARI mudel-PAM	ARI mudel	ARI PAM
0,2	0,620(0,020)	0,643(0,018)	0,642(0,018)	0,725(0,039)	0,877(0,024)	0,756(0,035)
0,4	0,325(0,023)	0,403(0,021)	0,407(0,019)	0,498(0,078)	0,626(0,037)	0,470(0,049)
0,6	0,114(0,016)	0,246(0,035)	0,289(0,016)	0,395(0,128)	0,315(0,049)	0,247(0,040)

Uurides ARI kriteeriumite keskmisi, näeme, et mudelipõhise klasteranalüüsi korral on väärtused suuremad, st mudelipõhise analüüsi korral saadud klasterdused on tegelike klassidega rohkem kooskõlas kui PAM-algoritmi klasterdused. Paneme tähele, et $\delta = 0{,}2$ korral on nii mudelipõhise kui ka PAM-algoritmi klasterduste keskmine kooskõla tegelike klassidega väga hea. Mudelipõhise analüüsi ja PAM-algoritmi klasterduste ARI keskmised lähevad kattuvuse suurenemisega väiksemaks, samal ajal standardhälbed lähevad suuremaks.

Märgime, et sagedaseim mudel, mis esines $\delta = 0{,}2$ korral 44 korda ning $\delta = 0{,}4$ ja $\delta = 0{,}6$ korral 51 korda, on kitsendustega segumudel $[\varepsilon_k^l]$, mille korral moodi tõenäosus sõltub nii klastrist kui ka tunnusest.

Tulemused komponentide võrdsete kaalude korral. Klasteranalüüsi tulemused võrdsete kaalude korral, $\pi_1 = \pi_2 = \pi_3 = {}^1/_3$, on ära toodud tabelis 9. Võrdsete kaalude korral näeme samuti, et klastrite kattuvuse suurenemisega kaasneb kriteeriumite keskmiste vähenemine. Antud olukorras on mudelipõhise analüüsi $\bar{s}_K$ keskmised natuke suuremad kui PAM-algoritmi korral, kuid nende põhjal ei saa kindlaks teha kumb meetod on antud olukorras sobilikum. Kuna PAM-algoritmi korral tekivad umbes sama suured klastrid, siis võrdsete kaalude korral võiks antud meetod teostada klasteranalüüsi paremini kui erinevate kaalude korral. Üldiselt aga seda ei ole näha ja saadud PAM-algoritmi tulemused erinevate kaalude korral on isegi natuke paremad. Tegelike klasside $\bar{s}_K$ keskmised võrdsete kaalude korral on jällegi mõlemast teisest $\bar{s}_K$ väärtusest väiksemad. Selline tulemus võib olla tingitud asjaolust, et andmete genereerimisel võib saada vaatlusi, mis ei ole tüüpilised antud segujaotuse komponendi korral ja on tüüpilisemad mingi teise komponendi suhtes.

Tabel 9. Klasteranalüüsi tulemuste keskmised ja standardhälbed 100 andmestiku ja võrdsete kaalude korral

δ	$\bar{s}_K$ tegelik	$\bar{s}_K$ mudel	$\bar{s}_K$ PAM	ARI mudel-PAM	ARI mudel	ARI PAM
0,2	0,620(0,022)	0,643(0,020)	0,640(0,020)	0,831(0,122)	0,804(0,030)	0,795(0,032)
0,4	0,323(0,021)	0,408(0,021)	0,401(0,020)	0,682(0,165)	0,531(0,036)	0,514(0,038)
0,6	0,116(0,014)	0,269(0,033)	0,264(0,017)	0,468(0,183)	0,246(0,040)	0,235(0,037)

Vaadeldes ARI kriteeriumite keskmisi, paneme tähele, et $\delta = 0{,}2$ ja $\delta = 0{,}4$ puhul on mudelipõhise analüüsi ja PAM-algoritmi kooskõla keskmised üldiselt head, kuid omavad

suurimaid standardhälbeid. Märgime, et mudelipõhise ja PAM-algoritmi klasterduste kooskõla on suurem, kui nende samade klasterduste ja tegelike klasside kooskõla. Kui võrrelda saadud klasterdusi tegelike klassidega, siis on mudelipõhise klasteranalüüsi ja tegelike klasside kooskõla minimaalselt suurem kui PAM-algoritmi korral, aga üldiselt ei saa öelda, et ARI kohaselt on mingi meetod sobilikum.

Märgime, et sagedaseim mudel, mis esines $\delta = 0{,}2$ korral 59 korda, $\delta = 0{,}4$ korral 75 korda ja $\delta = 0{,}6$ korral 55 korda, on kitsendustega segumudel $[\varepsilon^l]$, mille korral moodi tõenäosus sõltub ainult tunnusest.

Võrreldes nüüd tulemusi erinevate klastrite kattuvuste korral, saab öelda, et kui kattuvus on väiksem, $\delta = 0{,}2$, siis nii erinevate kui ka võrdsete kaalude korral annavad mõlemad meetodid häid tulemusi. Kui aga kattuvus suureneb, $\delta = 0{,}4$, $\delta = 0{,}6$, siis võrdsete kaalude korral ei ole meetodite vahel väga selget erinevust ja arvatavasti juba kattuvuse $\delta = 0{,}4$ korral ei ole võimalik klastreid hästi eraldada. Erinevate kaalude korral on aga mudelipõhine klasteranalüüs sobilikum ARI kohaselt. Üldiselt saab öelda, et kui soovitakse saada tegelikele klassidele võimalikult sarnaseid klastreid, siis annab mudelipõhine klasteranalüüs parema tulemuse antud näite kohaselt. Kuna artiklis Anderlucci ja Hennig (2014) on tehtud suurem analüüs erinevate olukordade põhjal, siis selle järeldused on usaldusväärsemad ja võib märgata, et paljud antud näite tulemused on kooskõlas artikli tulemustega. Paneme tähele, et samamoodi on suurema kattuvuse korral mudelipõhine analüüs sobilikum kui PAM-algoritm. Kui kattuvus on väiksem, siis tehakse artiklis järeldus, et kumbki meetod ei ole parem. Ainuke järeldus, mis ei tule artiklist välja, on meetodite samaväärsus võrdsete kaalude korral suvalise kattuvuse puhul. Põhjuseks võivad olla erinevad tunnuste väärtuste arvu kombinatsioonid ja kaalude erinevus.

Optimaalseim klastrite arv genereeritud andmestike korral. Kui klastrite kattuvus $\delta = 0{,}2$ on pigem väike ja sellise kattuvuse korral töötavad mõlemad meetodid hästi, valides optimaalseimaks klastrite arvuks tegelikele komponentide arvule vastava klastrite arvu, siis kattuvuste $\delta = 0{,}4$ ja $\delta = 0{,}6$ hakkavad meetodid teostama klasteranalüüsi juba erinevalt. Seega uurime optimaalset komponentide arvu mõlema meetodi korral suuremate kattuvuste korral, mis on ära toodud tabelites 10 ja 11.

Tabelis 10 näeme, et mudelipõhise klasteranalüüsi korral valitakse kahekomponendiline segumudel tihti juba kattuvuse $\delta = 0{,}4$ korral. Paneme tähele, et võrdsete kaalude puhul on

kolmekomponendiliste mudelite arv suurem kui erinevate kaalude puhul. Antud juhul ei teki olukorda, kus klasterdamist ei toimu. Samal ajal $\delta = 0{,}6$ korral on erinevate kaalude puhul parimad segumudelid ainult ühekomponendilised, st klastreid ei ole võimalik eraldada. Võrdsete kaalude korral on aga parimate mudelite hulgas ka kahekomponendilisi mudeleid. Seega võib öelda, et suurema kattuvuse korral on mudelipõhise klasteranalüüsi korral parimateks mudeliteks pigem väiksema klastrite arvuga mudelid. Antud tulemus selgitab ka ühe andmestiku põhjal saadud tulemuste väikseid kriteeriumite väärtusi kolmekomponendiliste klasterduste korral $\delta = 0{,}6$ puhul.

Tabel 10. Parima segumudeli klastrite arvu sagedustabel mudelipõhise klasteranalüüsi korral

	Erinevad kaalud			Võrdsed kaalud		
	k=1	k=2	k=3	k=1	k=2	k=3
$\delta = 0{,}4$ mudel	0	89	11	0	63	37
$\delta = 0{,}6$ mudel	100	0	0	77	23	0

Tabelis 11 näeme, et $\delta = 0{,}4$ korral on erinevate kaalude puhul kõik optimaalseimad klasterdused kahekomponendilised PAM-algoritmi korral. Kuid võrdsete kaalude korral lisanduvad optimaalsemate klasterduste hulka ka kolmekomponendilised klasterdused. Kui kattuvus suureneb, $\delta = 0{,}6$, siis nägime eelnevas tabelis, et klastreid ei olegi võimalik eraldada, seega PAM-algoritmi korral antud olukorda ei vaadelda.

Tabel 11. Optimaalseima klasterduse klastrite arvu sagedustabel PAM-algoritmi korral

	Erinevad kaalud		Võrdsed kaalud	
	k=2	k=3	k=2	k=3
$\delta = 0{,}4$ PAM	100	0	76	24

Kokkuvõtteks, mida suuremaks läheb klastrite kattuvus, seda väiksemaks läheb optimaalseima klasterduse komponentide arv ehk seda raskem on andmete genereerimiseks kasutatud komponente eraldada. Paneme tähele, et nii erinevate kui ka võrdsete kaalude korral klastrite kattuvuse $\delta = 0{,}4$ puhul eraldab mudelipõhine klasteranalüüs klastreid paremini.

Kokkuvõte

Käesoleva magistritöö eesmärk oli võrrelda kaht klasteranalüüsi meetodit, täpsemalt mudelipõhist klasteranalüüsi ja K-medoidide meetodit, millest üks põhineb mudelite hindamisel ja teine vaatluste omavahelistel kaugustel. Uuriti, kuidas käituvad need kaks meetodit erinevates olukordades, kas ja kuidas sõltub meetodite erinevus klastrite kattuvusest ja kui suur on „suur“ klastrite kattuvus. Mõlemad meetodid olid töö käigus põhjalikult kirjeldatud ja illustreeritud erinevate näidete abil. Samuti defineeriti meetodite võrdlemiseks vajalikud kriteeriumid, täpsemalt kohandatud Randi indeksi ja keskmise silueti laiuse kriteerium. Klasteranalüüsi teostamiseks genereeriti andmestikud, kasutades väiksemaid ja suuremaid klastrite kattuvuse määrasid ning erinevaid ja võrdseid segujaotuse komponentide kaalusid.

Esimeses väiksemas simulatsiooni ülesandes vaadeldi neli olukorda, kus kasutati väiksemat ja suuremat klastrite kattuvust ($\delta = 0{,}2,\ \delta = 0{,}6$) ja kahte segujaotuse komponentide kaalude kombinatsiooni ($\pi_1 = 0{,}15, \pi_2 = 0{,}35, \pi_3 = 0{,}5$ ning $\pi_1 = \pi_2 = \pi_3 = {}^1\!/_3$). Klasteranalüüs teostati klastrite arvu $K = 2{,}3{,}4$ jaoks. Tulemused näitasid, et kui kattuvus on $\delta = 0{,}2$, siis saavad mõlemad meetodid klastrite eraldamisega hästi hakkama mõlema komponentide kaalude komplekti korral. Paneme tähele, et keskmise silueti laiuse kriteeriumi väärtused olid mõlema meetodi puhul umbes samaväärsed. Kohandatud Randi indeksi kohaselt on väiksema klastrite arvuga mudelipõhise analüüsi ja PAM-algoritmi klasterdused rohkem kooskõlas kui suurema klastrite arvuga. Kui aga kattuvus on $\delta = 0{,}6$, siis on kohandatud Randi indeksi väärtused palju väiksemad, st meetodid teostavad klasteranalüüsi erinevalt.

Ühe andmestiku põhjal saadud klasteranalüüsi tulemuste abil on raske teha põhjalikke järeldusi, seega teostati klasteranalüüs näites 6 saja genereeritud andmestiku jaoks. Vaadeldi klastrite arvu $K = 3$ jaoks, kuna genereeritud vaatlused olid kirjeldatud kolmekomponendilise segujaotuse abil. Selles ülesandes vaadeldi lisaks kattuvust $\delta = 0{,}4$. Väiksema kattuvuse $\delta = 0{,}2$ korral töötavad nii mudelipõhine analüüs kui ka PAM-algoritm samaväärselt. Kui kattuvus suureneb, siis on tegelike klasside ja mudelipõhise analüüsi klasterdused rohkem kooskõlas kui PAM-algoritmi klasterdused, eriti erinevate komponentide kaalude korral. Seega kui klasteranalüüsi eesmärgiks on leida tegelikele klassidele võimalikult sarnased klastrid, siis on suurema kattuvuse korral sobilikum mudelipõhine klasteranalüüs. Sama järelduse saab teha ka segumudelite optimaalsemate klastrite arvu tulemuste kohaselt (tabelid 10 ja 11), mis näitasid, et mudelipõhise analüüsi korral on kolmekomponendiliste klasterduste hulk suurem kui PAM-

algoritmi korral. Üldiselt võib väita, et mida suurem on klastrite kattuvus, seda väiksem on optimaalsemate klasterduste klastrite arv, st seda raskem on klastreid üksteisest eraldada. Kui vaadelda kattuvust $\delta = 0{,}6$, siis klasterdamist ei toimu üldse, seega võib kattuvuse määra $\delta = 0{,}6$ suureks nimetada.

Antud magistritöö näitas, et klasteranalüüsi tulemused sõltuvad väga palju parameetrite valikust (komponentide kaalud, klastrite kattuvus, tunnuste arv jne) ja kasutatavast klasterdusmeetodist. Saadud informatsiooni põhjal saab antud teemat uurida edasi ja koostada uusi näiteid, mis annaksid lisatulemusi töös võrreldud klasteranalüüsi meetodite kohta.

Kasutatud kirjandus

Anderlucci, L., Hennig, C. (2014). The Clustering of Categorical Data: A Comparison of a Model-based and a Distance-based Approach. *Communications in Statistics – Theory and Methods*, 43:3, 704-721, doi:10.1080/03610926.2013.86665.

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

Hubert, L., Arabie, P. (1985). Comparing partitions. *Journal of Classification,* 2:193-218, doi: 10.1007/BF01908075.

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

Kaufman, L., Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New Jersey.

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* (Magistritöö). Tartu Ülikool.

Steinley, D., Hubert, L. (2008). Order-Constrained Solutions in K-Means Clustering: Even Better Than Being Globally Optimal. Psychometrika 73, 647, doi: 10.1007/s11336-008-9058-z.

Xu, R., Wunsch, D.C. (2008). *Clustering.* Wiley, New Jersey.

Lisa. Simulatsioonide tulemuste R-kood kattuvuse 0,6 korral

Funktsioonid *klasterda1, mixmod_tulem1, alfa* olid võetud magistritööst Mirski (2019).

```
library(cluster)
library(Rmixmod)
library(mclust)
library(nomclust)
genAndmed<-function(n,pi,delta){
  #Multinomiaalsete jaotuste tõenäosuste arvutamine
  alfa1=alfa2=alfa3=rep(NA,times=sum(m)) #K=3
  loendur=0
  for (j in 1:p){
    for (h in 1:m[j]){
      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)
      alfa3[loendur]=alfa(k=3,j=j,h=h,delta=delta)
    }
  }

  #Ühtlasest jaotusest genereeritud arvude abil leiame klastrite mahud
  #VÕRDSED KAALUD
  #set.seed(12)
  #uhtlane=runif(n)
  #n1=sum(uhtlane<pi[1])
  #n2=sum(pi[1]<uhtlane&uhtlane<2*pi[2])
  #n3=n-n1-n2

  #ERINEVAD KAALUD
  set.seed(12)
  uhtlane=runif(n)
  n1=sum(uhtlane<pi[1])
  n2=sum(uhtlane<pi[2])
  n3=n-n1-n2

  #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[3:4])))
  x1c=data.frame(t(rmultinom(n=n1,size=1,prob=alfa1[5:7])))
  x1d=data.frame(t(rmultinom(n=n1,size=1,prob=alfa1[8:10])))
  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[3:4])))
```

```
  x2c=data.frame(t(rmultinom(n=n2,size=1,prob=alfa2[5:7])))
  x2d=data.frame(t(rmultinom(n=n2,size=1,prob=alfa2[8:10])))
  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")

  #Kolmas klaster
  x3a=data.frame(t(rmultinom(n=n3,size=1,prob=alfa3[1:2])))
  x3b=data.frame(t(rmultinom(n=n3,size=1,prob=alfa3[3:4])))
  x3c=data.frame(t(rmultinom(n=n3,size=1,prob=alfa3[5:7])))
  x3d=data.frame(t(rmultinom(n=n3,size=1,prob=alfa3[8:10])))
  x3a3=data.frame("K3"=apply(x3a,1,function(x)  which(x==max(x))))
  x3b3=data.frame("K3"=apply(x3b,1,function(x)  which(x==max(x))))
  x3c3=data.frame("K3"=apply(x3c,1,function(x)  which(x==max(x))))
  x3d3=data.frame("K3"=apply(x3d,1,function(x)  which(x==max(x))))
  x3=cbind(x3a3,x3b3,x3c3,x3d3,3)
  colnames(x3)=c("X1","X2","X3","X4","K")

  #Kõik genereeritud vaatlused koos
  valim=rbind(x1,x2,x3)
  for (i in 1:ncol(valim))
    valim[,i]=as.factor(valim[,i])
  return(valim)
}

#MUDELIPÕHINE KLASTERDAMINE
klasterda1<-function(andmed){
  #mudelid [e], [e^j] ja [e_k^j]

mudelid=mixmodMultinomialModel(listModels=c("Binary_pk_E","Binary_pk_Ej","
Binary_pk_Ekj"))
  mixmod=mixmodCluster(data=data.frame(andmed),nbCluster=1:4,
                       dataType="qualitative",models=mudelid,

strategy=mixmodStrategy(nbTry=40,nbIterationInAlgo=1000,
epsilonInAlgo=0.0001),
                       seed=12,criterion="ICL")}

#Funktsioon,mis teeb Rmixmod tulemuste andmestiku
mixmod_tulem1<-function(valim,mudelid,n,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_bic
    tulemused[i,5]=ICLtapne(valim,mudelid[[i]],m=m,n) #ICL
  }
  colnames(tulemused)=c("Mudel","K","Toepara","ICL_bic","ICL")
  return(tulemused)}
```

```
#Funktsioon, mis arvutab täpse ICL kriteeriumi väärtuse
ICLtapne<-function(andmed,mudel,m,n){
  nk=table(mudel@partition) #klastrite mahud hat(n)_k,
  K=length(nk) #klastrite arv K
  #Tunnuste väärtuste sagedused klastrites (p=4)
  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(l
gamma(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(n)+K/2)+K*sum(lgamma(m/2)-m*lgamma(1/2))
  return(round(ICL,3))
}

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

#NÄIDE 21, erinevad kaalud, suur kattuvus
K=3; pi2=c(0.15,0.35, 0.5)
p=4; m=c(2,2,3,3)
n1=500
valim5<-genAndmed(n=n1,pi=pi2, delta=0.6)
#save(valim5, file="valim5.RDS")
#load("valim5.RDS")

#mudelipõhine klasteranalüüs, valim5
mudel5=klasterda1(andmed=valim5[,-5])
(tulemused5=mixmod_tulem1(valim=valim5,mudelid=mudel5@results,n=n1,m=m))
dist5<-sm(valim5[,-5])

#2 klastri
(parim5_2kl=mudel5["results"][[4]])
summary(silhouette(parim5_2kl@partition,dist5))
#plot(silhouette(parim5_2kl@partition,dist5))



#3 klastri
(parim5_3kl=mudel5["results"][[9]])
table(valim5$K,parim5_3kl@partition)
summary(silhouette(parim5_3kl@partition,dist5))
```

```
#plot(silhouette(parim5_3kl@partition,dist5))



#4 klastri
(parim5_4kl=mudel5["results"][[12]])
summary(silhouette(parim5_4kl@partition,dist5))
#plot(silhouette(parim5_4kl@partition,dist5),main="mudel_4kl_0.6")


#K-medoids, valim5
#2 klastri
pam_fit5_2kl<-pam(dist5, diss = TRUE,k=2)
summary(pam_fit5_2kl)$silinfo$avg.width
#plot(pam_fit5_2kl,main="PAM_2kl_0.6")


#3 klastri
pam_fit5_3kl<-pam(dist5, diss = TRUE,k=3)
summary(pam_fit5_3kl)$silinfo$avg.width
table(valim5$K, summary(pam_fit5_3kl)$clustering)
#plot(pam_fit5_3kl,main="PAM_3kl_0.6")


#4 klastri
pam_fit5_4kl<-pam(dist5, diss = TRUE,k=4)
summary(pam_fit5_4kl)$silinfo$avg.width
#plot(pam_fit5_4kl,main="PAM_4kl_0.6")


#võdleme meetodeid 2 klastri korral
adjustedRandIndex(x=parim5_2kl@partition, y=summary(pam_fit5_2kl)$clusteri
ng)

#võdleme meetodeid 3 klastri korral
adjustedRandIndex(x=parim5_3kl@partition, y=summary(pam_fit5_3kl)$clusteri
ng)

#võdleme meetodeid 4 klastri korral
adjustedRandIndex(x=parim5_4kl@partition, y=summary(pam_fit5_4kl)$clusteri
ng)

#NÄIDE 22, samad kaalud, suur kattuvus
K=3; pi1=c(1/3,1/3, 1/3)
p=4; m=c(2,2,3,3)
n1=500
valim7<-genAndmed(n=n1,pi=pi1, delta=0.6)
#save(valim7, file="valim7.RDS")
#load("valim7.RDS")

#mudelipõhine klasteranalüüs, valim7
mudel7=klasterda1(andmed=valim7[,-5])
(tulemused7=mixmod_tulem1(valim=valim7,mudelid=mudel7@results,n=n1,m=m))
dist7<-sm(valim7[,-5])
```

```
#2 klastri
(parim7_2kl=mudel7["results"][[4]])
summary(silhouette(parim7_2kl@partition,dist7))
#plot(silhouette(parim7_2kl@partition,dist7),main="mudel_2kl_0.6")


#3 klastri
(parim7_3kl=mudel7["results"][[8]])
table(valim7$K,parim7_3kl@partition)
summary(silhouette(parim7_3kl@partition,dist7))
plot(silhouette(parim7_3kl@partition,dist7),main="Klasterdus kolmekomponen
dilise segumudeli korral")
abline(v=0.29, col="red")


#4 klastri
(parim7_4kl=mudel7["results"][[12]])
summary(silhouette(parim7_4kl@partition,dist7))
#plot(silhouette(parim7_4kl@partition,dist7),main="mudel_4kl_0.6")


#K-medoids, valim7
#2 klastri
pam_fit7_2kl<-pam(dist7, diss = TRUE,k=2)
summary(pam_fit7_2kl)$silinfo$avg.width
#plot(pam_fit7_2kl,main="PAM_2kl_0.6")


#3 klastri
pam_fit7_3kl<-pam(dist7, diss = TRUE,k=3)
table(valim7$K, summary(pam_fit7_3kl)$clustering)
summary(pam_fit7_3kl)$silinfo$avg.width
#plot(pam_fit7_3kl,main="PAM_3kl_0.6")


#4 klastri
pam_fit7_4kl<-pam(dist7, diss = TRUE,k=4)
summary(pam_fit7_4kl)$silinfo$avg.width
#plot(pam_fit7_4kl,main="PAM_4kl_0.6")


#võdleme meetodeid 2 klastri korral
adjustedRandIndex(x=parim7_2kl@partition, y=summary(pam_fit7_2kl)$clusteri
ng)

#võdleme meetodeid 3 klastri korral
adjustedRandIndex(x=parim7_3kl@partition, y=summary(pam_fit7_3kl)$clusteri
ng)

#võdleme meetodeid 4 klastri korral
adjustedRandIndex(x=parim7_4kl@partition, y=summary(pam_fit7_4kl)$clusteri
ng)
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Anastassia Ugrjumova,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Mudelipõhise klasteranalüüsi ja K-medoidide meetodi võrdlemine kvalitatiivsete tunnustega andmete klasterdamisel“, 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.

Anastassia Ugrjumova

18.06.2020