Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Sören Mirski

Mudelipõhine klasteranalüüs

Matemaatika ja statistika õppekava

Matemaatilise statistika eriala

Magistritöö (30 EAP)

Juhendaja: vanemteadur Kristi Kuljus

Tartu 2019

Mudelipõhine klasteranalüüs

Magistritöö

Sören Mirski

Lühikokkuvõte. Mudelipõhiste klasterdamismeetodite korral eeldatakse, et vaatlusi on sobiv kirjeldada segujaotuse abil, mille iga komponent määrab ühe klastri. Mudelipõhine klasteranalüüs leiab üha enam kasutamist, kuna sel juhul asendub sobiva klasterdamismeetodi valik statistilise mudeli valikuga ja optimaalse klastrite arvu leidmise ülesanne taandub segujaotuse komponentide arvu hindamise ülesandeks. Käesoleva magistritöö eesmärk on anda ülevaade mudelipõhise klasteranalüüsi teostamisest kvantitatiivsete, kvalitatiivsete ning segatüüpi tunnuste korral. Töö esimeses peatükis defineeritakse segujaotused erinevat tüüpi tunnuste korral ning selgitatakse, kuidas EM-algoritmiga nende jaotuste parameetreid hinnatakse. Lisaks tuletatakse niinimetatud integreeritud klassifitseerimistõepära ehk ICL kriteerium, mida mudelipõhise klasteranalüüsi korral kasutatakse segumudeli sobivuse ja klastrite arvu hindamiseks. Töö teises peatükis rakendatakse mudelipõhist klasterdamist Tartu Ülikooli Eesti Geenivaramu biomarkerite andmestikule, mis sisaldab nii kvantitatiivseid kui ka kvalitatiivseid tunnuseid.

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

Märksõnad: klasteranalüüs, tõenäosusjaotused, normaaljaotus, simulatsioon, geenidoonorid, biomarkerid, R (programmeerimiskeel).

Model-Based Clustering

Master's thesis

Sören Mirski

Abstract. Clustering procedures based on mixture models assume that observations can be described by a mixture of distributions where each component corresponds to one cluster. Model-based clustering has become popular because then the problem of choosing an appropriate clustering method can be recast as the choice of a suitable statistical model. Also, the task of finding the number of clusters in the data is replaced with estimating the suitable number of components in the mixture model. The purpose of this master's thesis is to provide an overview of model-based clustering in the case of quantitative, qualitative or mixed-type variables. In the first chapter mixture models used for different types of data are defined and the process of estimating the model parameters with the EM algorithm is explained. The integrated completed likelihood (ICL) criterion, which is used to assess the suitability of the mixture model and the number of clusters, is also derived. In the second part of the thesis model-based clustering is applied to a mixed-type biomarker data set from the Estonian Genome Center at the University of Tartu.

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

Keywords: cluster analysis, probability distributions, normal distribution, simulation, gene donors, biomarkers, R (programming language).

Sisukord

Sissejuhatus

Järgnev klasteranalüüsi arutelu põhineb raamatul Hennig, Meila, Murtagh ja Rocci (2016), kui ei ole märgitud teisiti. Klasteranalüüs on üks populaarsemaid juhendamata õppe ja andmeanalüüsi meetodeid, mis leiab laialdast kasutust masinõppes, geneetikas, pildianalüüsis ja paljudes muudes valdkondades. Klasteranalüüsi eesmärk on grupeerida vaatlused neil mõõdetud tunnuste alusel nii, et ühte gruppi paigutatud vaatlused oleksid võimalikult sarnased. Peamised klasteranalüüsi meetodid võib jagada kahte klassi: vaatlustevahelistel kaugustel põhinevad ning mudelipõhised meetodid. Kõige tavalisemad kaugusi kasutavad meetodid on hierarhiline klasterdamine ja K-keskmiste meetod, mille tulemused sõltuvad suurel määral sellest, kas ja kuidas tunnuseid enne meetodi rakendamist teisendatakse või standardiseeritakse. Hierarhilise klasteranalüüsi tulemusi mõjutab ka valitud kaugusmõõt ja klastrite ühendamise meetod. Mudelipõhise klasteranalüüsi korral eeldatakse, et klasterdatavad vaatlused pärinevad mitmemõõtmelisest segujaotusest, mille parameetrid tuleb hinnata.

Klasteranalüüsi kui statistilise meetodi põhiprobleem on range klastri definitsiooni puudumine. Teatud määral on probleemiks ka lai valik aastate jooksul välja töötatud erinevaid lähenemisi ja meetodeid. Seetõttu peaks iga klasteranalüüsi ülesanne algama võimalikult detailse eesmärgi püstitamisega, millele järgneb vajadustest ning soovidest lähtuvalt sobiva klasterdamismeetodi valik. Kaugustel põhinevate meetoditega võrreldes on mudelipõhise lähenemise eelisteks, et sobiva meetodi valik asendub statistilise mudeli valikuga ning klastrite arvu leidmise ülesanne taandub segujaotuse komponentide arvu hindamise ülesandeks, sest hinnatava segujaotuse iga komponent määrab ühe klastri.

Klasteranalüüsi teostamine on raskendatud, kui vaatluseid kirjeldavad tunnused on segatüüpi (nii kvantitatiivsed kui ka kvalitatiivsed). Sama tüüpi tunnuste korral on andmete klasterdamiseks välja töötatud palju meetodeid, kuid segatüüpi tunnuste korral sobivaid meetodeid on vähe (Foss ja Markatou, 2018, lk 3). Seetõttu teisendatakse segatüüpi andmestike korral tunnuseid sageli nii, et ühte tüüpi tunnuste jaoks mõeldud meetodid oleksid rakendatavad. Näiteks kui suurem osa mõõdetud tunnustest on kvantitatiivsed, saab üksikud kvalitatiivsed tunnused asendada nende võimalikele väärtustele vastavate indikaatortunnustega. Sel juhul on probleemiks aga lisanduvate tunnuste arv ning nende sobiv kaalumine. Väikeste kaalude korral ei pruugi kvalitatiivsed tunnused vaatluste grupeerimisel üldse abiks olla ja suurte kaalude korral võib nende üksikute tunnuste mõju klasteranalüüsi tulemustele olla teiste tunnustega võrreldes liiga suur. Seega on segatüüpi tunnustega andmete klasterdamisel põhiprobleemiks mõlemat tüüpi tunnuste tasakaalustatud kasutamine. Mudelipõhise klasteranalüüsi korral ei ole kvantitatiivsete

ja kvalitatiivsete tunnuste samaaegne käsitlemine keeruline, kuna eeldatakse, et nende tunnuste rühmad on iga klastri sees sõltumatud. See võimaldab rakendada mõlemale rühmale sobivat tõenäosusmudelit.

Antud magistritöö eesmärk on anda ülevaade mudelipõhise klasteranalüüsi teostamisest kvantitatiivsete, kvalitatiivsete ning segatüüpi tunnuste korral. Töö esimeses peatükis defineerime sõltuvalt tunnuste tüübist eeldatavad segujaotused ning selgitame, kuidas EM-algoritmiga nende jaotuste parameetreid hinnatakse. Lisaks tuletame niinimetatud integreeritud klassifitseerimistõepära ehk ICL kriteeriumi, mida mudelipõhise klasteranalüüsi korral kasutatakse erinevate kitsenduste ning komponentide arvuga hinnatud segumudelite hulgast parima mudeli valimiseks. Mudelipõhise klasteranalüüsi käitumise illustreerimiseks erineva kujuga andmete korral viime läbi kaks simulatsiooninäidet. Esimene osa lõpeb rakendustarkvara R lisapaketi „Rmixmod“ kirjeldusega. Seda paketti kasutame antud töös mudelipõhise klasteranalüüsi teostamiseks. Töö teises peatükis rakendame mudelipõhist klasteranalüüsi Tartu Ülikooli Eesti Geenivaramust pärinevale segatüüpi tunnustega andmestikule ja interpreteerime klasterdamise tulemusi.

Töö kirjutamiseks on kasutatud tekstitöötlusprogrammi LaTeX ning klasteranalüüside läbiviimiseks rakendustarkvara R versiooni 3.4.1 ja lisapaketi „Rmixmod“ versiooni 2.1.2.

Autor tänab professor Krista Fischerit abi eest andmete muretsemisel töö praktilise osa teostamiseks ning juhendajat Kristi Kuljust asjakohaste nõuannete, paranduste ja kasulike ideede eest.

1 Mudelipõhine klasteranalüüs

Käesolevas peatükis kirjeldame, kuidas teostada mudelipõhist klasteranalüüsi, kui klasterdatavaid vaatlusi iseloomustavad tunnused on kvantitatiivsed, kvalitatiivsed või segatüüpi. Allikana on kasutatud raamatut Hennig jt (2016), kui ei ole viidatud teisiti.

1.1 Kvantitatiivsete andmete klasterdamine

Tähistagu $\mathbf{x}_1, \ldots, \mathbf{x}_n$ klasterdatavaid vaatluseid, millel on mõõdetud p kvantitatiivse tunnuse väärtused. Eesmärk on paigutada vaatlused $\mathbf{x}_1, \ldots, \mathbf{x}_n$ nende omavahelise sarnasuse alusel paarikaupa lõikumatutesse hulkadesse ehk klastritesse $C_1, \ldots, C_K$. Mudelipõhise klasteranalüüsi korral eeldatakse, et vaatluste kirjeldamiseks sobib mitmemõõtmeline segujaotus.

Definitsioon 1. 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 $\mathbf{X}$ on juhuslike komponentide $\mathbf{X}_1, \ldots, \mathbf{X}_K$ segu, kui selle tihedusfunktsioon avaldub kujul

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

kus $\mathbf{x} \in \mathbb{R}^p$, $\pi_1, \ldots, \pi_K$ on komponentide kaalud ($\pi_k \geq 0$ ja $\sum_{k=1}^{K} \pi_k = 1$), f_k on komponendi $\mathbf{X}_k$ 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 (1) kõigi parameetrite hulka.

Teisisõnu eeldatakse, et vaatlused $\mathbf{x}_1, \ldots, \mathbf{x}_n$ on tihedusfunktsiooniga (1) juhusliku vektori $\mathbf{X}$ sõltumatud realisatsioonid. Olgu $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$ vaatlusele $\mathbf{x}_i$ vastav indikaatorvektor, kus $z_{ik} = 1$, kui vaatluse $\mathbf{x}_i$ genereeris segujaotuse (1) komponent $\mathbf{X}_k$, ning $z_{ik} = 0$ vastasel juhul, $k = 1, \ldots, K$, $i = 1, \ldots, n$. Olgu $\mathbf{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ vektoritele $\mathbf{z}_1, \ldots, \mathbf{z}_n$ vastavate juhuslike vektorite hulk, kus iga vektor $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iK})'$ on multinomiaalse jaotusega $\mathrm{Mult}_K(1, \boldsymbol{\pi})$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$. Sel juhul vastab segujaotuse igale komponendile üks klaster. Et tundmatu vektor $\mathbf{z}_i$ näitab, millisesse gruppi vaatlus $\mathbf{x}_i$ kuulub, seisnebki mudelipõhine klasteranalüüs sisuliselt indikaatorvektorite $\mathbf{z}_1, \ldots, \mathbf{z}_n$ hindamises.

Vaatluste $\mathbf{x}_1, \ldots, \mathbf{x}_n$ klastritesse $C_1, \ldots, C_K$ määramiseks kasutatakse tinglike tõenäosuste

$$\gamma_k(\mathbf{x}_i; \boldsymbol{\theta}) = \mathbb{P}\{Z_{ik} = 1 \,|\, \mathbf{X} = \mathbf{x}_i\} = \frac{\mathbb{P}\{Z_{ik} = 1\} f(\mathbf{x}_i; \boldsymbol{\theta} \,|\, Z_{ik} = 1)}{\sum_{j=1}^{K} \mathbb{P}\{Z_{ij} = 1\} f(\mathbf{x}_i; \boldsymbol{\theta} \,|\, Z_{ij} = 1)} = \frac{\pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)}{f(\mathbf{x}_i; \boldsymbol{\theta})} \tag{2}$$

hinnanguid. Suurus $\gamma_k(\mathbf{x}_i; \boldsymbol{\theta})$ on tinglik tõenäosus, et vaatlus $\mathbf{x}_i$ on parameetritega $\boldsymbol{\theta}$

segujaotuse k-nda komponendi realisatsioon, mistõttu tõlgendatakse seda sageli ka kui „vastutust“ (*responsibility*), mille komponent $\mathbf{X}_k$ võtab vaatluse $\mathbf{x}_i$ kirjeldamisel. Need tõenäosused defineerivad vaatluste niinimetatud pehme klasterduse (*soft clustering*), kus igale vaatlusele on ühe klastrisse kuuluvust näitava väärtuse asemel vastavusse seatud kõikidesse klastritesse kuulumise tõenäosused. Enamasti on klasteranalüüsi korral eesmärk aga määrata iga vaatlus parajasti ühte klastrisse. Mudelipõhise klasteranalüüsi korral on loomulik paigutada iga vaatlus klastrisse, mille korral tingliku tõenäosuse (2) hinnang on kõige suurem. Seega vaatlus $\mathbf{x}_i$ paigutatakse klastrisse C_k, kui $\hat{z}_{ik} = 1$, kus

$$\hat{z}_{ik} = \begin{cases} 1, & \text{kui } k = \arg\max_h \hat{\gamma}_h(\mathbf{x}_i; \boldsymbol{\theta}), \\ 0, & \text{vastasel juhul,} \end{cases} \tag{3}$$

ja $\hat{\gamma}_k(\mathbf{x}_i; \boldsymbol{\theta})$ on tõenäosuse (2) hinnang. Tulemuseks on paarikaupa lõikumatud klastrid, mis on määratud hulkadega $C_k = \{i \,|\, \hat{z}_{ik} = 1\}$, $k = 1, \ldots, K$. Need hulgad sisaldavad vastavatesse klastritesse paigutatud vaatluste indekseid.

1.1.1 Normaaljaotuste segu

Kvantitatiivsete andmete mudelipõhisel klasterdamisel on kõige tavalisem vaadelda mitmemõõtmelise normaaljaotusega komponentide segu, sest praktikas moodustavad vaatlused sageli elliptilise kujuga klastreid, mida on mugav normaaljaotuste abil modelleerida. Lisaks on normaaljaotuste korral segujaotuse (1) parameetrite hindamine EM-algoritmiga lihtne, sest hinnangute valemid on analüütiliselt leitavad. Kui segujaotuse kõik komponendid on p-mõõtmelise normaaljaotusega, siis komponendi $\mathbf{X}_k$ jaotuse parameetriteks on keskväärtusvektor $\boldsymbol{\mu}_k \in \mathbb{R}^p$ ning kovariatsioonimaatriks $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p \times p}$ ehk $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k = 1, \ldots, K$. Komponentide tihedused on kujul

$$f_k(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\}, \quad \mathbf{x} \in \mathbb{R}^p. \tag{4}$$

Normaaljaotuste segu korral on klastrid elliptilise kujuga, kus ellipsoidide keskpunktideks on keskväärtusvektorid $\boldsymbol{\mu}_k$ ja ellipsoidide kuju (pooltelgede pikkuse ning suuna) määravad kovariatsioonimaatriksid $\boldsymbol{\Sigma}_k$, $k = 1, \ldots, K$.

Kui maatriksitele $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K$ ei ole seatud ühtegi kitsendust (näiteks ei eeldata, et $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ või $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}_p$, $\sigma^2 > 0$, $\forall k$), siis segujaotuse parameetrite arv võib olenevalt vaadeldavast tunnuste ja komponentide arvust olla üsna kõrge, sest iga kovariatsioonimaatriks sisaldab $\frac{1}{2}p(p+1)$ tundmatut parameetrit. Artiklis Banfield ja Raftery (1993) soovitatakse erinevate kitsenduste seadmiseks lähtuda dekompositsioonist

$$\boldsymbol{\Sigma}_k = \lambda_k D_k A_k D_k', \tag{5}$$

kus $\lambda_k = |\mathbf{\Sigma}_k|^{1/p}$, D_k on $\mathbf{\Sigma}_k$ omavektorite ortogonaalne maatriks ning A_k on diagonaalmaatriks, mille peadiagonaalil on kahanevalt järjestatud maatriksi $\mathbf{\Sigma}_k$ normaliseeritud omaväärtused: $A_k = \text{diag}(\lambda_{k1}/\lambda_k, \ldots, \lambda_{kp}/\lambda_k)$, $\lambda_{k1} \geq \ldots \geq \lambda_{kp} > 0$, $k = 1, \ldots, K$. Üldistatud dispersioon $|\mathbf{\Sigma}_k| = \prod_{j=1}^{p} \lambda_{kj}$ kirjeldab normaaljaotuse $\mathcal{N}_p(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$ hajuvust. Näiteks kui kahemõõtmelisel juhul on mõlema tunnuse dispersioon väike ehk jaotuse tõenäosusmass on kontsentreeritud keskväärtusvektori $\boldsymbol{\mu}_k$ ümber, siis $|\mathbf{\Sigma}_k| = \lambda_{k1}\lambda_{k2}$ on väike, sest omaväärtused λ_{k1} ja λ_{k2} on väikesed. Kui hajuvust ühes või mõlemas suunas suurendada, suureneb ka üldistatud dispersioon. Seega määrab tegur λ_k segujaotuse k-ndale komponendile vastava ellipsoidi ehk klastri suuruse. Maatriks D_k määrab selle klastri suuna ja maatriks A_k selle kuju (kas klaster on pigem sfääriline või elliptiline).

Näide 1. Kovariatsioonimaatriksite esitus suuruse, suuna ja kuju parameetrite abil
Vaatleme järgmist kolme kovariatsioonimaatriksit:

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 3 & 4 \\ 4 & 9 \end{pmatrix}, \quad \mathbf{\Sigma}_2 = \frac{1}{\sqrt{11}} \begin{pmatrix} 9{,}79 & 1{,}98 \\ 1{,}98 & 12{,}76 \end{pmatrix}, \quad \mathbf{\Sigma}_3 = \frac{1}{\sqrt{11}} \begin{pmatrix} 19{,}58 & 3{,}96 \\ 3{,}96 & 25{,}52 \end{pmatrix}.$$

Nende maatriksite dekompositsiooni (5) tegurid on kujul $\lambda_1 = \lambda_2 = \sqrt{11}$, $\lambda_3 = 2\sqrt{11}$,

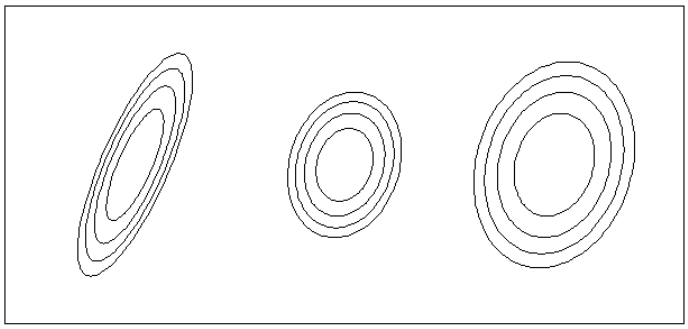
$$D_1 = D_2 = D_3 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}$$

ja

$$A_1 = \frac{1}{\sqrt{11}} \begin{pmatrix} 11 & 0 \\ 0 & 1 \end{pmatrix}, \quad A_2 = A_3 = \begin{pmatrix} 1{,}25 & 0 \\ 0 & 0{,}8 \end{pmatrix}.$$

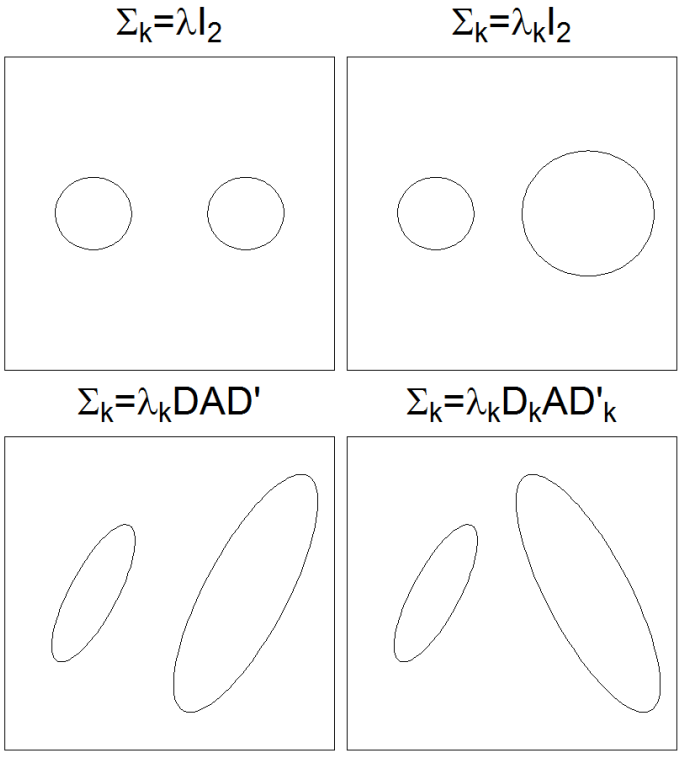
Järelikult kovariatsioonimaatriksitega $\mathbf{\Sigma}_1$ ja $\mathbf{\Sigma}_2$ normaaljaotuste hajuvusellipsid (tihedusfunktsiooni ristlõiked) on sama suuruse ja suunaga, aga erineva kujuga (vt joonis 1 lk 10). Kuna maatriksi A_1 diagonaali esimene element on 11 korda suurem kui teine element, on kovariatsioonimaatriksiga $\mathbf{\Sigma}_1$ normaaljaotuse hajuvusellipsid pikliku kujuga. Maatriksi A_2 diagonaalelemendid on aga samas suurusjärgus, mistõttu joonisel 1 lk 10 keskmise jaotuse hajuvusellipsid on sfäärilisemad (võrreldes kovariatsioonimaatriksiga $\mathbf{\Sigma}_1$ on maatriksis $\mathbf{\Sigma}_2$ tunnuste dispersioonide erinevus väiksem ning tunnustevaheline kovariatsioon on samuti väiksem). Jooniselt 1 lk 10 on lisaks selgesti näha, et kovariatsioonimaatriksitega $\mathbf{\Sigma}_2$ ja $\mathbf{\Sigma}_3$ normaaljaotuste hajuvusellipsid on sama suuna ja kujuga, kuid nende suurus on erinev ehk nad on proportsionaalsed.

Lubades osadel kovariatsioonimaatriksi esituse (5) teguritel segujaotuse komponentide vahel varieeruda ja fikseerides ülejäänud tegurid, jõutakse erinevate kitsendustega segumudelite klassideni. Joonisel 2 lk 10 on näitena kujutatud kahemõõtmeliste kovariatsioonimaatriksite $\mathbf{\Sigma}_1$ ja $\mathbf{\Sigma}_2$ hajuvusellipsid nelja erineva kitsenduste kombinatsiooni korral.



Joonis 1. Kovariatsioonimaatriksitega $\mathbf{\Sigma}_1$ (vasakul), $\mathbf{\Sigma}_2$ (keskel) ja $\mathbf{\Sigma}_3$ (paremal) normaaljaotuste hajuvusellipsid

Esimesel alamjoonisel on eeldatud sfäärilisi (diagonaalseid ja ühise dispersiooniga) ning võrdseid kovariatsioonimaatrikseid, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \lambda \boldsymbol{I}_2$. Seetõttu on saadavad kaks klastrit ringikujulised ja sama suured. Teisel juhul on segujaotuse komponentidele vastavad kaks klastrit samuti ringikujulised, kuid erineva suurusega, sest λ_1 ja λ_2, mis antud juhul on tunnuste dispersioonid klastrites, pole võrdsed. Kolmandal alamjoonisel kujutatud ellipsid on sama suuna ($D_1 = D_2 = D$) ja kujuga ($A_1 = A_2 = A$), kuid erineva suurusega ehk nad on proportsionaalsed. Neljanda kitsenduste kombinatsiooni korral on ellipsitel erinev suurus ja suund, aga sama kuju (samuti proportsionaalsed).



Joonis 2. Kovariatsioonimaatriksitega $\mathbf{\Sigma}_k = \lambda_k D_k A_k D_k'$ $(k = 1, 2)$ normaaljaotuste hajuvusellipsid nelja erineva kitsenduste kombinatsiooni korral

Jooniste 1 ja 2 tegemiseks kirjutatud tarkvara R kood on toodud lisas 1.

1.1.2 Segujaotuse parameetrite hindamine EM-algoritmiga

Segujaotuse (1) parameetrite hinnangute leidmiseks kasutatakse EM-algoritmi (*expectation-maximization algorithm*), mis on suurima tõepära meetodi iteratiivne modifikatsioon latentsete tunnuste korral. Antud alapeatükis kirjeldame EM-algoritmi kahte üldist sammu ning toome ära algoritmi ja parameetrite hinnangute valemid normaaljaotuste segu korral. Allikana on kasutatud raamatut Bishop (2006), kui ei ole märgitud teisiti.

Olgu $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ kõigi vaatluste hulk ja $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ indikaatorvektorite hulk, mille iga element näitab, millise segujaotuse komponendi realisatsioon vastav vaatlus on. Tähistagu $\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ vaatluste $\{\mathbf{x}, \mathbf{z}\}$ logaritmilist tõepärafunktsiooni, kus $\boldsymbol{\theta}$ on segujaotuse kõigi parameetrite hulk. Kuna vektorid $\mathbf{z}_1, \ldots, \mathbf{z}_n$ on tegelikult tundmatud, ei ole funktsiooni $\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ maksimeerimine praktikas võimalik. Kui maksimeerida vaatluste $\mathbf{x}$ logaritmilist tõepärafunktsiooni $\ln p(\mathbf{x}; \boldsymbol{\theta})$, sisaldavad leitud parameetrite hinnangud kaudselt vektoreid $\mathbf{z}$. Selle probleemi lahendamiseks kasutatakse vektorite $\mathbf{z}$ kohta teada-olevat informatsiooni, mis piirdub tingliku jaotusega $p(\mathbf{z} \,|\, \mathbf{x}; \boldsymbol{\theta})$. Funktsiooni $\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ asemel vaadeldakse tinglikku keskväärtust $E\left[\ln p(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta}) \,|\, \mathbf{x}\right]$, kus $\mathbf{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ on indikaatoritele $\mathbf{z}_1, \ldots, \mathbf{z}_n$ vastavate multinomiaalse jaotusega juhuslike vektorite hulk. Keskväärtuse $E\left[\ln p(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta}) \,|\, \mathbf{x}\right]$ arvutamine on EM-algoritmi keskväärtustamise samm ehk niinimetatud E-samm. Järgneb algoritmi maksimeerimise samm ehk niinimetatud M-samm, kus leitakse parameetrid, mis maksimeerivad E-sammul saadud tingliku keskväärtuse. Kui $\boldsymbol{\theta}^{\text{vana}}$ tähistab parameetrite esialgseid hinnanguid, siis kirjeldatud kahe sammu teostamise tulemuseks on uued hinnangud $\boldsymbol{\theta}^{\text{uus}}$. Seejuures on garanteeritud, et logaritmilise tõepärafunktsiooni $\ln p(\mathbf{x}; \boldsymbol{\theta})$ väärtus kasvab või jääb samaks ehk ei kahane: $\ln p(\mathbf{x}; \boldsymbol{\theta}^{\text{uus}}) \geq \ln p(\mathbf{x}; \boldsymbol{\theta}^{\text{vana}})$.

Kokkuvõttes on EM-algoritm järgmine:

1. Fikseeri segujaotuse parameetrite algväärtused $\boldsymbol{\theta}^{\text{vana}}$ ja arvuta logaritmiline tõepära $\ln p(\mathbf{x}; \boldsymbol{\theta}^{\text{vana}})$.

2. **E-samm**: Leia tundmatute indikaatorvektorite tinglik jaotus $p(\mathbf{z} \,|\, \mathbf{x}; \boldsymbol{\theta}^{\text{vana}})$ ning funktsiooni $\ln p(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta})$ tinglik keskväärtus

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{\text{vana}}) = E\left[\ln p(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta}) \,|\, \mathbf{x}; \boldsymbol{\theta}^{\text{vana}}\right] = \sum_{\mathbf{z}} \ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) p(\mathbf{z} \,|\, \mathbf{x}; \boldsymbol{\theta}^{\text{vana}}).$$

3. **M-samm**: Leia parameetrite uued hinnangud $\boldsymbol{\theta}^{\text{uus}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{vana}})$.

4. Arvuta $\ln p(\mathbf{x}; \boldsymbol{\theta}^{\text{uus}})$ ning kontrolli algoritmi koondumist logaritmilise tõepärafunktsiooni $\ln p(\mathbf{x}; \boldsymbol{\theta})$ või parameetrite hinnangute koondumise kaudu. Kui

algoritm ei ole koondunud, siis võta $\boldsymbol{\theta}^{\text{vana}} \leftarrow \boldsymbol{\theta}^{\text{uus}}$ ja mine tagasi sammu 2 juurde.

Praktikas on tavalisem EM-algoritmi koondumise kontrollimiseks kasutada parameetrite hinnangute koondumise asemel logaritmilise tõepärafunktsiooni $\ln p(\mathbf{x};\boldsymbol{\theta})$ koondumist. Levinud on kaks koondumiskriteeriumi. Esimese kriteeriumi korral kontrollitakse, kas $\ln p(\mathbf{x};\boldsymbol{\theta}^{\text{uus}}) - \ln p(\mathbf{x};\boldsymbol{\theta}^{\text{vana}}) < \varepsilon$ mingi väikse väärtuse $\varepsilon > 0$ korral. Teisel juhul algoritmi iteratsiooniprotsess lõpetatakse, kui logaritmilise tõepära suhteline muutus on väiksem kui fikseeritud konstant ε ehk

$$\frac{\ln p(\mathbf{x};\boldsymbol{\theta}^{\text{uus}}) - \ln p(\mathbf{x};\boldsymbol{\theta}^{\text{vana}})}{|\ln p(\mathbf{x};\boldsymbol{\theta}^{\text{vana}})|} < \varepsilon. \tag{6}$$

Kuna funktsiooni $\ln p(\mathbf{x};\boldsymbol{\theta})$ suurusjärk sõltub vaatluste arvust (mida rohkem on vaatlusi, seda suurem on logaritmilise tõepära absoluutväärtus), on esimese kriteeriumi kasutamine raskendatud, sest konstandi ε sobiva väärtuse valimine on keeruline. Seetõttu on tarkvara R paketis „Rmixmod“, mida antud töös kasutame mudelipõhise klasteranalüüsi teostamiseks, implementeeritud vaid suhtelise muutuse kriteerium (Langrognet jt, 2018). Kuna EM-algoritm koondub enamasti logaritmilise tõepärafunktsiooni $\ln p(\mathbf{x};\boldsymbol{\theta})$ lokaalseks maksimumiks, tuleb heade hinnangute leidmiseks algoritmi rakendada mitu korda erinevaid parameetrite alglähendeid kasutades.

Segujaotuse (1) parameetrite hindamisel EM-algoritmiga tuleb meeles pidada, et komponentide kaalud $\pi_1, \ldots, \pi_K$ summeeruvad üheks. Et ka algoritmi M-sammul leitavad uued kaalud $\pi_1^{\text{uus}}, \ldots, \pi_K^{\text{uus}}$ seda kitsendust rahuldaksid, saab kasutada Lagrange'i kordajate meetodit. Sel juhul maksimeeritakse tingliku keskväärtuse $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{\text{vana}})$ asemel funktsiooni $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{\text{vana}}) + \lambda(\sum_{k=1}^{K}\pi_k - 1)$, kus lisaparameeter λ võimaldab maksimeerimisülesande lahendamisel nimetatud kitsendust arvesse võtta.

Olgu nüüd vaatluse all segujaotus, mille iga komponent $\mathbf{X}_k$ on p-mõõtmelise normaaljaotusega $\mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ juhuslik vektor, $k = 1, \ldots, K$. Siis

$$\ln p(\mathbf{x};\boldsymbol{\theta}) = \ln\left[\prod_{i=1}^{n} f(\mathbf{x}_i;\boldsymbol{\theta})\right] = \sum_{i=1}^{n} \ln\left[\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right],$$

kus tihedusfunktsioon f_k on antud valemiga (4). Vaatluste $\{\mathbf{x},\mathbf{z}\}$ logaritmiline tõepärafunktsioon on sel juhul kujul

$$\ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) = \ln\left\{\prod_{i=1}^{n}\prod_{k=1}^{K}[\pi_k f_k(\mathbf{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)]^{z_{ik}}\right\} = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left[\ln\pi_k + \ln f_k(\mathbf{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right], \tag{7}$$

sest $(\mathbf{x}_1,\mathbf{z}_1), \ldots, (\mathbf{x}_n,\mathbf{z}_n)$ on sõltumatud ja $p(\mathbf{x}_i,\mathbf{z}_i) = p(\mathbf{x}_i \mid \mathbf{z}_i)p(\mathbf{z}_i)$ iga $i = 1, \ldots, n$ korral. Tähistagu $\boldsymbol{\theta}^{\text{vana}}$ olemasolevate parameetrite hinnangute hulka. Avaldisest (7) on ilmne, et tingliku keskväärtuse $E\left[\ln p(\mathbf{x},\mathbf{Z};\boldsymbol{\theta}) \mid \mathbf{x};\boldsymbol{\theta}^{\text{vana}}\right]$ arvutamiseks on vaja leida indikaatorile z_{ik}

vastava juhusliku suuruse Z_{ik} tinglik keskväärtus, mis avaldub kujul

$$E(Z_{ik} \,|\, \mathbf{x}; \boldsymbol{\theta}^{\text{vana}}) = \mathbb{P}\{Z_{ik} = 1 \,|\, \mathbf{x}; \boldsymbol{\theta}^{\text{vana}}\} = \frac{\pi_k^{\text{vana}} f_k(\mathbf{x}_i; \boldsymbol{\mu}_k^{\text{vana}}, \boldsymbol{\Sigma}_k^{\text{vana}})}{f(\mathbf{x}_i; \boldsymbol{\theta}^{\text{vana}})}$$
$$= \frac{\pi_k^{\text{vana}} f_k(\mathbf{x}_i; \boldsymbol{\mu}_k^{\text{vana}}, \boldsymbol{\Sigma}_k^{\text{vana}})}{\sum_{j=1}^{K} \pi_j^{\text{vana}} f_j(\mathbf{x}_i; \boldsymbol{\mu}_j^{\text{vana}}, \boldsymbol{\Sigma}_j^{\text{vana}})} = \gamma_k(\mathbf{x}_i; \boldsymbol{\theta}^{\text{vana}}).$$

See on tinglik tõenäosus, et vaatlus $\mathbf{x}_i$ on parameetritega $\boldsymbol{\theta}^{\text{vana}}$ normaaljaotuste segu k-nda komponendi realisatsioon. Järelikult

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{\text{vana}}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_k(\mathbf{x}_i; \boldsymbol{\theta}^{\text{vana}}) \left[\ln \pi_k + \ln f_k(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right].$$

Seega saab normaaljaotuste segu korral EM-algoritmi kirja panna järgmisel kujul:

1. Fikseeri parameetrite algväärtused $\boldsymbol{\theta}^{\text{vana}}$ ehk $\boldsymbol{\mu}_k^{\text{vana}}$, $\boldsymbol{\Sigma}_k^{\text{vana}}$ ja π_k^{vana}, $k = 1, \ldots, K$, ning arvuta logaritmiline tõepära $\ln p(\mathbf{x}; \boldsymbol{\theta}^{\text{vana}})$.

2. **E-samm**: Arvuta tinglikud tõenäosused

$$\gamma_k(\mathbf{x}_i; \boldsymbol{\theta}^{\text{vana}}) = \frac{\pi_k^{\text{vana}} f_k(\mathbf{x}_i; \boldsymbol{\mu}_k^{\text{vana}}, \boldsymbol{\Sigma}_k^{\text{vana}})}{\sum_{j=1}^{K} \pi_j^{\text{vana}} f_j(\mathbf{x}_i; \boldsymbol{\mu}_j^{\text{vana}}, \boldsymbol{\Sigma}_j^{\text{vana}})}, \quad k = 1, \ldots, K, \; i = 1, \ldots, n.$$

3. **M-samm**: Leia parameetrite uued hinnangud $\boldsymbol{\theta}^{\text{uus}}$:

$$\boldsymbol{\mu}_k^{\text{uus}} = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_k(\mathbf{x}_i; \boldsymbol{\theta}^{\text{vana}}) \mathbf{x}_i,$$

$$\boldsymbol{\Sigma}_k^{\text{uus}} = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_k(\mathbf{x}_i; \boldsymbol{\theta}^{\text{vana}}) (\mathbf{x}_i - \boldsymbol{\mu}_k^{\text{uus}})(\mathbf{x}_i - \boldsymbol{\mu}_k^{\text{uus}})',$$

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

   kus $n_k = \sum_{i=1}^{n} \gamma_k(\mathbf{x}_i; \boldsymbol{\theta}^{\text{vana}})$, $k = 1, \ldots, K$.

4. Arvuta $\ln p(\mathbf{x}; \boldsymbol{\theta}^{\text{uus}})$ ning kontrolli algoritmi koondumist logaritmilise tõepärafunktsiooni $\ln p(\mathbf{x}; \boldsymbol{\theta})$ koondumise kaudu. Kui algoritm ei ole koondunud, siis võta $\boldsymbol{\theta}^{\text{vana}} \leftarrow \boldsymbol{\theta}^{\text{uus}}$ ja mine tagasi sammu 2 juurde.

Kuna EM-algoritm võib üsna aeglaselt koonduda, on parameetrite heade alglähendite kasutamine sageli suureks abiks. Näiteks normaaljaotuste segu korral kasutatakse keskväärtusvektorite $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ alglähenditena tihti arvutuslikult kiire K-keskmiste meetodiga leitud klastrite keskpunkte. Sel juhul maatriksite $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K$ algväärtusteks võetakse saadud klastrite põhjal arvutatud kovariatsioonimaatriksid ning komponentide kaalud võrdsustatakse klastrite osakaaludega. Tihti valitakse algoritmi algväärtused aga juhuslikult.

1.1.3 Segumudeli sobivuse ja klastrite arvu hindamine

Mudelipõhise klasteranalüüsi korral asendub sobiva klasterdamismeetodi valik statistilise mudeli valikuga ning klastrite arvu leidmise ülesanne taandub segujaotuse komponentide arvu hindamise ülesandeks. Erinevate kitsenduste ja komponentide arvuga segumudelite hindamisel tekib aga küsimus, kuidas nende hulgast parim valida. Kuna segujaotuste parameetrite hinnangud leitakse tõepära $L(\boldsymbol{\theta}) = p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\theta})$ maksimeerimise kaudu, saab sobiva segumudeli valimiseks kasutada erinevaid informatsioonikriteeriume.

Definitsioon 2. Akaike informatsioonikriteerium avaldub kujul (Akaike, 1974)

$$\mathrm{AIC} = -2 \ln L(\hat{\boldsymbol{\theta}}) + 2\nu, \tag{8}$$

kus $L(\hat{\boldsymbol{\theta}})$ on mudeli maksimeeritud tõepära ja ν on mudeli parameetrite arv.

Valemis (8) olev karistusliige 2ν takistab Akaike kriteeriumi väärtuse vähenemist, kui mudeli keerukuse ehk parameetrite arvu suurendamise tulemusena logaritmilise tõepärafunktsiooni väärtus aina kasvab. Kuna AIC karistusliige ei arvesta mudeli parameetrite hindamiseks kasutatavate vaatluste arvuga, defineeris Schwarz (1978) alternatiivina niinimetatud Bayesi kriteeriumi, mille karistusliige sõltub nii mudeli parameetrite arvust kui ka valimimahust.

Definitsioon 3. Bayesi informatsioonikriteerium avaldub kujul

$$\mathrm{BIC} = -2 \ln L(\hat{\boldsymbol{\theta}}) + \nu \ln n, \tag{9}$$

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

Valem (9) erineb artiklis Schwarz (1978) toodud avaldisest, kuna seal defineeritakse Akaike kriteerium valemiga $\mathrm{AIC} = \ln L(\hat{\boldsymbol{\theta}}) - \nu$ ning analoogselt $\mathrm{BIC} = \ln L(\hat{\boldsymbol{\theta}}) - \frac{\nu}{2} \ln n$. Antud töös defineerime Bayesi kriteeriumi valemiga (9), kuna see on kooskõlas AIC ajaloolise definitsiooniga (8) ja see valem on kasutusel ka tarkvara R lisapaketis „Rmixmod“ (Lebret jt, 2015, lk 10). Mõlema defineeritud informatsioonikriteeriumi kohaselt on kõige parem mudel, mille kriteeriumi väärtus on minimaalne.

Mudelipõhise klasteranalüüsi korral on täheldatud, et kui tegelik mudel vaatluse all olevate segumudelite hulka ei kuulu, siis BIC eelistab tegelikust suurema komponentide arvuga mudeleid. Selle probleemi lahendamiseks töötasid Biernacki, Celeux ja Govaert (2000) välja niinimetatud integreeritud klassifitseerimistõepära (edaspidi ICL) kriteeriumi, mis võrreldes Bayesi kriteeriumiga võtab lisaks veel arvesse, et vaatluse all olevad segumudelid on hinnatud klasteranalüüsi teostamise eesmärgil. Järgnev ICL kriteeriumi tuletuskäik

põhineb eelnimetatud artiklil.

Vaatluste $\{\mathbf{x}, \mathbf{z}\} = \{(\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n)\}$ integreeritud tõepärafunktsioon avaldub kujul

$$p(\mathbf{x}, \mathbf{z}) = \int_{\boldsymbol{\Theta}} p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{10}$$

kus

$$p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} [\pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)]^{z_{ik}},$$

$\boldsymbol{\Theta}$ on segujaotuse (1) parameeterruum ja $\pi(\boldsymbol{\theta})$ on selle segujaotuse kõigi parameetrite $\boldsymbol{\theta}$ eeljaotus. Tõepära (10) logaritmi lähendamiseks kasutatakse valemit

$$\ln p(\mathbf{x}, \mathbf{z}) \approx \ln p(\mathbf{x}, \mathbf{z}; \hat{\boldsymbol{\theta}}^*) - \frac{\nu}{2} \ln n, \tag{11}$$

kus $\hat{\boldsymbol{\theta}}^* = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ ning ν on vaatluse all oleva K-komponendilise segumudeli parameetrite arv. Kuna vektorid $\mathbf{z}_1, \ldots, \mathbf{z}_n$ on tundmatud, pole suurima tõepära hinnang $\hat{\boldsymbol{\theta}}^*$ leitav, kuid piisavalt suure valimimahu korral saab selle asemel kasutada hinnangut $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}; \boldsymbol{\theta})$, kus $p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\theta})$. Lõpuks, asendades tundmatud klastrite indikaatorid $\mathbf{z}_1, \ldots, \mathbf{z}_n$ hinnangutega (3), jõutaksegi niinimetatud ICL kriteeriumini

$$\mathrm{ICL} = -2 \ln p(\mathbf{x}, \hat{\mathbf{z}}; \hat{\boldsymbol{\theta}}) + \nu \ln n. \tag{12}$$

Kuigi lähendusel (11) puudub range teoreetiline põhjendus, näidati artiklis Biernacki jt (2000) mitmete simulatsioonide ja praktiliste ülesannete abil, et ICL kriteerium on sobiv meetod segumudeli kuju ja komponentide arvu valimiseks mudelipõhise klasteranalüüsi korral. Selle kriteeriumi korral on hinnatud segumudelite ehk kandidaatmudelite hulgast samuti parim minimaalse ICL väärtusega mudel. Sarnaselt Bayesi kriteeriumiga erineb ka valem (12) artiklis Biernacki jt (2000) esitatud avaldisest. Kasutame avaldist (12), et defineeritud kolme kriteeriumi valemid oleksid kooskõlas (väiksem väärtus on parem).

Järgnevas arutelus kasutame avaldiste lihtsustamiseks tähistuste $f(\mathbf{x}_i; \boldsymbol{\theta})$ ja $\gamma_k(\mathbf{x}_i; \boldsymbol{\theta})$ asemel tähistusi $f(\mathbf{x}_i)$ ja $\gamma_k(\mathbf{x}_i)$. Kuna

$$p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} (\gamma_k(\mathbf{x}_i))^{z_{ik}} \left[\frac{\pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)}{\gamma_k(\mathbf{x}_i)} \right]^{z_{ik}} = \prod_{i=1}^{n} \prod_{k=1}^{K} (f(\mathbf{x}_i))^{z_{ik}} (\gamma_k(\mathbf{x}_i))^{z_{ik}},$$

siis

$$\begin{aligned} \ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln f(\mathbf{x}_i) + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln \gamma_k(\mathbf{x}_i) \\ &= \sum_{i=1}^{n} \ln f(\mathbf{x}_i) + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln \gamma_k(\mathbf{x}_i), \end{aligned}$$

sest $\sum_{k=1}^{K} z_{ik} = 1$ iga $i = 1, \ldots, n$ korral. Seega

$$\ln p(\mathbf{x}, \hat{\mathbf{z}}; \hat{\boldsymbol{\theta}}) = \ln p(\mathbf{x}; \hat{\boldsymbol{\theta}}) + \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik} \ln \hat{\gamma}_k(\mathbf{x}_i)$$

ja ICL kriteeriumi avaldise (12) saab kirja panna järgmiselt:

$$\begin{aligned} \mathrm{ICL} &= -2 \ln L(\hat{\boldsymbol{\theta}}) + \nu \ln n - 2 \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik} \ln \hat{\gamma}_k(\mathbf{x}_i) \\ &= \mathrm{BIC} - 2 \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik} \ln \hat{\gamma}_k(\mathbf{x}_i). \end{aligned}$$

Järelikult on ICL kriteerium võrdne Bayesi kriteeriumiga, millele on liidetud kahekordne karistusliige $-\sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik} \ln \hat{\gamma}_k(\mathbf{x}_i) \geq 0$, mis iseloomustab hinnatud segumudeli võimet vaatluseid $\mathbf{x}_1, \ldots, \mathbf{x}_n$ klasterdada. Kuna $E(Z_{ik} \,|\, \mathbf{x}) = \mathbb{P}\{Z_{ik} = 1 \,|\, \mathbf{x}\} = \gamma_k(\mathbf{x}_i)$, siis kasutatakse karistusliikmena ka vaatluste pehme klasterduse määrava maatriksi $(\hat{\gamma}_k(\mathbf{x}_i))$ entroopiat $-\sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\gamma}_k(\mathbf{x}_i) \ln \hat{\gamma}_k(\mathbf{x}_i) \geq 0$. Sel juhul on ICL kriteerium kujul

$$\mathrm{ICL} = \mathrm{BIC} - 2 \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\gamma}_k(\mathbf{x}_i) \ln \hat{\gamma}_k(\mathbf{x}_i). \tag{13}$$

Kui hinnatud segumudeli komponendid on üksteisest hästi eraldatud, määravad tinglikud tõenäosused $\hat{\gamma}_k(\mathbf{x}_i)$ selge vaatluste $\mathbf{x}_1, \ldots, \mathbf{x}_n$ klasterduse ning karistusliige $-\sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\gamma}_k(\mathbf{x}_i) \ln \hat{\gamma}_k(\mathbf{x}_i)$ on ligikaudu null. Kui segumudeli komponendid kattuvad suurel määral, näiteks komponentide ehk klastrite liiga suure arvu tõttu, on karistusliikme väärtus suur ja seetõttu on ka ICL suurema väärtusega. Klastrite eristamine on kõige raskem, kui iga vaatluse korral $\hat{\gamma}_k(\mathbf{x}_i) = 1/K$, $k = 1, \ldots, K$. Siis karistusliikme väärtus on $n \ln K$ (suurim võimalik väärtus). Järelikult pooldab ICL kriteerium segumudeleid, mille komponentide poolt määratud klastrite kattuvus on väike.

1.2 Kvalitatiivsete andmete klasterdamine

Olgu klasterdatavaid vaatlusi $\mathbf{x}_1, \ldots, \mathbf{x}_n$ kirjeldavad p tunnust kvalitatiivsed ja olgu neil tunnustel vastavalt $m_1, \ldots, m_p$ võimalikku väärtust. Vaatlusel $\mathbf{x}_i$ mõõdetud j-nda tunnuse väärtuse saab asendada vektoriga $(x_i^{j1}, \ldots, x_i^{jm_j})'$, kus $x_i^{jh} = 1$, kui sellel tunnusel on h-s võimalik väärtus, ja $x_i^{jh} = 0$ vastasel juhul, $j = 1, \ldots, p$, $i = 1, \ldots, n$. Seega on iga vaatlus $\mathbf{x}_i$ esitatav binaarse vektorina

$$\mathbf{x}_i = (x_i^{11}, \ldots, x_i^{1m_1}; x_i^{21}, \ldots, x_i^{2m_2}; \ldots; x_i^{p1}, \ldots, x_i^{pm_p})'.$$

Mudelipõhise klasteranalüüsi teostamiseks kvalitatiivsete tunnuste korral eeldatakse, et iga vaatluse $\mathbf{x} = (x^{11}, \ldots, x^{1m_1}; \ldots; x^{p1}, \ldots, x^{pm_p})'$ kirjeldamiseks sobib K-komponendiline „mitmemõõtmeliste“ multinomiaalsete jaotuste segu. Selle segujaotuse tõenäosusfunktsioon on defineeritud järgmiselt:

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{M}_k(\mathbf{x}; \boldsymbol{\alpha}_k) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x^{jh}}, \tag{14}$$

kus $\pi_1, \ldots, \pi_K$ on komponentide kaalud ($\pi_k \geq 0$ ja $\sum_{k=1}^{K} \pi_k = 1$), α_k^{jh} on tõenäosus, et j-ndal tunnusel on h-s võimalik väärtus, kui $\mathbf{x}$ on selle segujaotuse k-nda komponendi realisatsioon, $\boldsymbol{\alpha}_k = (\alpha_k^{11}, \ldots, \alpha_k^{1m_1}; \ldots; \alpha_k^{p1}, \ldots, \alpha_k^{pm_p})'$ ning $\boldsymbol{\theta}$ tähistab segujaotuse kõigi parameetrite hulka.

Segujaotuses (14) on eeldatud, et mõõdetud p kvalitatiivset tunnust on iga komponendi sees sõltumatud, seda eeldust nimetatakse ka lokaalseks sõltumatuseks. Sellise tingliku sõltumatuse eelduse tõttu sisaldab tõenäosusfunktsiooni (14) iga komponent p multinomiaalse jaotuse korrutist, mille parameetrite vektorid on $\boldsymbol{\alpha}_k^j = (\alpha_k^{j1}, \ldots, \alpha_k^{jm_j})'$, kus $\sum_{h=1}^{m_j} \alpha_k^{jh} = 1$, $j = 1, \ldots, p$, $k = 1, \ldots, K$. Eespool nimetasime iga sellist korrutist mitmemõõtmeliseks multinomiaalseks jaotuseks ning kasutasime tähistust $\mathcal{M}_k(\mathbf{x}; \boldsymbol{\alpha}_k)$. Peamine põhjus niinimetatud lokaalse sõltumatuse eelduse tegemiseks on, et kvalitatiivsete tunnuste vahelise sõltuvuse iseloomustamiseks pole lihtsat näitajat nagu seda on kovariatsioon kahe kvantitatiivse tunnuse vahel. Näiteks normaaljaotuste segu korral modelleerivad komponentide kovariatsioonimaatriksid kvantitatiivsete tunnuste vahelist sõltuvusstruktuuri, kuid kvalitatiivsete tunnuste korral ei ole see võimalik.

Parameetrite hindamine

Multinomiaalsete jaotuste segu (14) parameetrite hindamiseks kasutatakse samuti EM-algoritmi. Olgu $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ vaatluste $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ klastritesse kuuluvust näitavate indikaatorvektorite hulk ja olgu $\mathbf{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ vastavate juhuslike vektorite hulk. Analoogselt normaaljaotuste seguga arvutatakse algoritmi E-sammul parameetrite väärtuseid $\boldsymbol{\theta}^{\text{vana}}$ kasutades tinglikud tõenäosused $\mathbb{P}\{Z_{ik} = 1 \,|\, \mathbf{X} = \mathbf{x}_i\}$,

$$\gamma_k(\mathbf{x}_i; \boldsymbol{\theta}^{\text{vana}}) = \frac{\pi_k^{\text{vana}} \mathcal{M}_k(\mathbf{x}_i; \boldsymbol{\alpha}_k^{\text{vana}})}{\sum_{l=1}^{K} \pi_l^{\text{vana}} \mathcal{M}_l(\mathbf{x}_i; \boldsymbol{\alpha}_l^{\text{vana}})}, \quad k = 1, \ldots, K,\ i = 1, \ldots, n.$$

Antud juhul avaldub vaatluste $\{\mathbf{x}, \mathbf{z}\}$ logaritmiline tõepärafunktsioon valemiga

$$\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left(\ln \pi_k + \sum_{j=1}^{p} \sum_{h=1}^{m_j} x_i^{jh} \ln \alpha_k^{jh} \right).$$

Funktsiooni $\ln p(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta})$ tinglik keskväärtus $E\left[\ln p(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta}) \,|\, \mathbf{x}; \boldsymbol{\theta}^{\text{vana}}\right]$ on kujul

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{vana}}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_k(\mathbf{x}_i; \boldsymbol{\theta}^{\text{vana}}) \left(\ln \pi_k + \sum_{j=1}^{p} \sum_{h=1}^{m_j} x_i^{jh} \ln \alpha_k^{jh} \right),$$

sest $E(Z_{ik} \,|\, \mathbf{x}; \boldsymbol{\theta}^{\text{vana}}) = \gamma_k(\mathbf{x}_i; \boldsymbol{\theta}^{\text{vana}})$. Segujaotuse (14) komponentide kaalud ja iga multinomiaalse jaotuse parameetrid summeeruvad üheks. Nende kitsendustega arvestamiseks parameetrite hinnangute leidmisel saab kasutada Lagrange'i kordajate meetodit. Seetõttu maksimeeritakse EM-algoritmi M-sammul tingliku keskväärtuse $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{vana}})$ asemel funktsiooni

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{vana}}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) + \sum_{k=1}^{K} \sum_{j=1}^{p} \lambda_{kj} \left(\sum_{h=1}^{m_j} \alpha_k^{jh} - 1 \right),$$

kus lisaparameetrid λ ja $\lambda_{11}, \ldots, \lambda_{Kp}$ garanteerivad, et parameetrite uued hinnangud rahuldavad eelnimetatud kitsendusi. Saadavad hinnangud $\boldsymbol{\theta}^{\text{uus}}$ avalduvad kujul

$$\pi_k^{\text{uus}} = \frac{n_k}{n}, \quad (\alpha_k^{jh})^{\text{uus}} = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_k(\mathbf{x}_i; \boldsymbol{\theta}^{\text{vana}}) x_i^{jh},$$

kus $n_k = \sum_{i=1}^{n} \gamma_k(\mathbf{x}_i; \boldsymbol{\theta}^{\text{vana}})$, $j = 1, \ldots, p$, $h = 1, \ldots, m_j$, $k = 1, \ldots, K$.

Kitsenduste seadmine

Kui klasterdatavate vaatluste arv on võrreldes mõõdetud tunnuste võimalike väärtuste arvuga liiga väike, ei pruugi segujaotuse (14) parameetrid hinnatavad olla või. Kui hinnatavate parameetrite arv on võrreldes vaatluste arvuga väga suur, ei pruugi hinnangud olla ka eriti usaldusväärsed. Kitsendusteta segumudelis on tundmatute parameetrite arv $(K-1) + K \sum_{j=1}^{p} (m_j - 1)$. Näiteks $K = 4$, $p = 7$, $m_1 = \ldots = m_7 = 5$ korral on vaja leida 115 parameetri hinnangud. Parameetrite arvu vähendamiseks saab ka segujaotusele (14) kitsendusi seada. Järgmised kaks lõiku põhinevad artiklil Lebret jt (2015).

Multinomiaalsete jaotuste segule kitsenduste seadmisel eeldatakse, et igas klastris on kõigil tunnustel üks väärtus teistest suurema tõenäosusega ehk igal tunnusel on üheselt määratud mood. Ülejäänud tõenäosusmass jagatakse tunnuse teiste väärtuste vahel võrdselt. Siis iga klastri k ja iga tunnuse j korral on tõenäosused $\boldsymbol{\alpha}_k^j = (\alpha_k^{j1}, \ldots, \alpha_k^{jm_j})'$ kujul $(\beta_k^j, \ldots, \beta_k^j, \gamma_k^j, \beta_k^j, \ldots, \beta_k^j)'$, kus $\gamma_k^j > \beta_k^j$ ning $\beta_k^j = (1 - \gamma_k^j)/(m_j - 1)$. Tähistagu $h(k, j) \in \{1, \ldots, m_j\}$ tõenäosuse γ_k^j positsiooni vektoris $\boldsymbol{\alpha}_k^j$. Kirjeldatud kitsenduste lihtsamaks tõlgendamiseks kasutatakse järgmist parameetrite esitust. Iga tõenäosuste vektor $\boldsymbol{\alpha}_k^j$ asendatakse parameetritega $\mathbf{a}_k^j = (a_k^{j1}, \ldots, a_k^{jm_j})'$, kus $a_k^{jh} = 1$, kui $h = h(k, j)$, ja $a_k^{jh} = 0$ vastasel juhul, ning $\varepsilon_k^j = 1 - \gamma_k^j$. Näiteks kui j-ndal tunnusel on neli

võimalikku väärtust, mille tõenäosused k-ndas klastris on $\boldsymbol{\alpha}_k^j = (0{,}2; 0{,}2; 0{,}4; 0{,}2)'$, siis $\mathbf{a}_k^j = (0, 0, 1, 0)'$ ja $\varepsilon_k^j = 0{,}6$, sest $\gamma_k^j = 0{,}4$.

Kirjeldatud kitsenduste korral avalduvad multinomiaalsete jaotuste tõenäosused nende uute parameetrite kaudu järgmiselt:

$$\alpha_k^{jh} = \begin{cases} 1 - \varepsilon_k^j, & \text{kui } h = h(k, j), \\ \varepsilon_k^j/(m_j - 1), & \text{vastasel juhul.} \end{cases}$$

Selliste parameetritega segumudelit tähistatakse $[\varepsilon_k^j]$, sest iga tunnuse korral on antud klastris tarvis hinnata üks parameeter, moodi tõenäosus $\gamma_k^j = 1 - \varepsilon_k^j$. Selle tõenäosuse positsiooni esialgses parameetrite vektoris $\boldsymbol{\alpha}_k^j$ ehk j-nda tunnuse moodi k-ndas klastris määrab vektor $\mathbf{a}_k^j$. Antud alapunkti alguses toodud näites langeb parameetrite arv kitsenduste $[\varepsilon_k^j]$ korral 31 parameetrile (rohkem kui kolmekordne vähenemine). Veel on võimalik defineerida kolm kitsendustega segumudelit. Esiteks, kui eeldada, et antud klastris on iga tunnuse moodi tõenäosus võrdne ehk $\gamma_k^j = \gamma_k$ iga $j = 1, \ldots, p$ korral, siis vastavat mudelit tähistatakse $[\varepsilon_k]$. Teiseks, kui j-nda tunnuse moodi tõenäosus on kõigis klastrites sama ehk $\gamma_k^j = \gamma^j$ iga $k = 1, \ldots, K$ korral, siis tulemuseks on mudel $[\varepsilon^j]$. Lõpuks, kui eeldada, et moodi tõenäosus γ_k^j ei sõltu vaadeldavast klastrist ega tunnusest, saadakse mudel $[\varepsilon]$. Kitsendusteta segumudelit (14) tähistatakse $[\varepsilon_k^{jh}]$, mis binaarsete tunnuste ($m_j = 2$, $j = 1, \ldots, p$) korral taandub mudeliks $[\varepsilon_k^j]$. Kokkuvõttes on defineeritud parametrisatsiooni korral vaatluse all viis mudelit: $[\varepsilon_k^{jh}]$, $[\varepsilon_k^j]$, $[\varepsilon_k]$, $[\varepsilon^j]$ ja $[\varepsilon]$.

Näide 2. Kvalitatiivsete andmete klasterdamine klastrite erineva kattuvuse korral
Kvalitatiivsete andmete mudelipõhisel klasterdamisel määravad tunnuste väärtuste tõenäosused ehk tunnuste jaotused klastrite kattuvuse määra. Kui mõõdetud tunnuse iga väärtus esineb suure tõenäosusega parajasti ühes grupis ja seda iga tunnuse korral, on saadavad klastrid üksteisest selgesti eraldatud ehk nende kattuvus on väike. Aga kui mõned väärtused esinevad sarnase tõenäosusega mitmes grupis, on klastrite kattuvus suurem. Käesoleva näite eesmärk on illustreerida mudelipõhise klasteranalüüsi käitumist erineva klastrite kattuvusega kvalitatiivsete andmete korral. Selleks genereerime kahe situatsiooni kohaselt $n = 1000$ ja $n = 3000$ vaatlust kahe komponendiga segujaotusest (14). Mõlemas situatsioonis kirjeldavad klasterdatavaid vaatlusi $p = 4$ kvalitatiivset tunnust, millel on vastavalt $m_1 = m_2 = 3$ ja $m_3 = m_4 = 4$ võimalikku väärtust. Vaadeldava segujaotuse komponentide kaalud olgu $\pi_1 = 0{,}3$ ja $\pi_2 = 0{,}7$ ehk vaatluste klasterdamisel on maksimaalne veamäär 30% (kui kõik vaatlused paigutatakse ühte klastrisse).

Erineva klastrite kattuvuse määraga andmete genereerimiseks lähtume järgmisest segu-

jaotuse parameetrite valemist (Biernacki, Celeux ja Govaert, 2010, lk 2996):

$$\alpha_k^{jh} = \begin{cases} \frac{1}{m_j} + (1-\delta)\frac{m_j-1}{m_j}, & \text{kui } h = [(k-1)\ mod\ m_j] + 1, \\ \frac{\left[1-\frac{1}{m_j}-(1-\delta)\frac{m_j-1}{m_j}\right]}{m_j-1}, & \text{vastasel juhul,} \end{cases} \tag{15}$$

kus konstandiga $\delta \in [0,1]$ saab reguleerida klastrite kattuvust ja *mod* tähistab jäägiga jagamist, $j = 1,\ldots,p$, $h = 1,\ldots,m_j$, $k = 1,\ldots,K$. Valemis (15) kasutatakse jäägiga jagamist kõige sagedasema väärtuse ehk moodi määramiseks. Segujaotuse komponentide parameetrite vektorid $\delta = 0$ korral on siin näites kujul

$$\boldsymbol{\alpha}_1 = (1,0,0;\quad 1,0,0;\quad 1,0,0,0;\quad 1,0,0,0)',$$
$$\boldsymbol{\alpha}_2 = (0,1,0;\quad 0,1,0;\quad 0,1,0,0;\quad 0,1,0,0)'.$$

Seega vastab $\delta = 0$ olukorrale, kus klastrid ei kattu, kõigil neljal tunnusel on mõlemas klastris erinevad väärtused. Klastrite kattuvus on maksimaalne $\delta = 1$ korral, sest siis on iga tunnuse kõik väärtused mõlemas klastris võrdvõimalikud ehk parameetrite vektorid on kujul

$$\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \left(\frac{1}{3},\frac{1}{3},\frac{1}{3};\quad \frac{1}{3},\frac{1}{3},\frac{1}{3};\quad \frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4};\quad \frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right)'.$$

Valemiga (15) arvutatud parameetritega segumudel vastab kitsendustega mudelile $[\varepsilon^j]$, sest iga tunnuse moodi tõenäosus on kõigis klastrites sama, $\gamma_k^j = \gamma^j$ iga $k = 1,\ldots,K$ korral. Ainult moodiks olevad väärtused on klastrites erinevad. Moodid on määratud vektoritega $\mathbf{a}_k^j = (a_k^{j1},\ldots,a_k^{jm_j})'$, kus $a_k^{jh} = 1$, kui $h = [(k-1)\ mod\ m_j] + 1$, ja $a_k^{jh} = 0$ vastasel juhul.

Mudelipõhise klasteranalüüsi teostamiseks kasutame tarkvara R lisapaketti „Rmixmod“, millest anname ülevaate alapeatükis 1.5. Mõlemas situatsioonis teostame klasteranalüüsi klastrite arvudega $K = 1,\ldots,4$. Parameetrite hinnangud leiame EM-algoritmiga, mida rakendame 20 korda erinevate alglähenditega (maksimaalne iteratsioonide arv on 1000 ning logaritmilise tõepärafunktsiooni suhtelise muutuse koondumiskriteeriumis (6) on $\varepsilon = 0{,}001$). Kuna antud näites vastavad genereeritud andmed segumudelile $[\varepsilon^j]$, jätame kitsendusteta mudeli $[\varepsilon_k^{jh}]$ vaatluse alt välja. Mudeli $[\varepsilon_k]$, kus moodi tõenäosus ei sõltu tunnusest, jätame ka vaatluse alt välja, kuna tõenäosuste (15) arvutamiseks kasutatakse tunnuste võimalike väärtuste arve, mis on erinevad. Seega jääb alles kolm segumudelit: $[\varepsilon_k^j]$ (moodi tõenäosus sõltub klastrist ja tunnusest), $[\varepsilon^j]$ (moodi tõenäosus sõltub ainult tunnusest) ning $[\varepsilon]$ (moodi tõenäosus ei sõltu klastrist ega tunnusest).

Kvalitatiivsete tunnuste korral ei ole vaja kasutada ICL kriteeriumi lähendust (12).

Kriteeriumi väärtuse saab täpselt välja arvutada valemiga (Biernacki jt, 2010, lk 2994)

$$\begin{aligned}\mathrm{ICL}_{\mathcal{M}} = &\sum_{k=1}^{K} \ln\Gamma\left(\hat{n}_k + \tfrac{1}{2}\right) + \sum_{k=1}^{K}\sum_{j=1}^{p}\left[\sum_{h=1}^{m_j} \ln\Gamma\left(\hat{u}_k^{jh} + \tfrac{1}{2}\right) - \ln\Gamma\left(\hat{n}_k + \tfrac{m_j}{2}\right)\right] \\ &+ \ln\Gamma\left(\tfrac{K}{2}\right) - K\ln\Gamma\left(\tfrac{1}{2}\right) - \ln\Gamma\left(n + \tfrac{K}{2}\right) \\ &+ K\sum_{j=1}^{p}\left[\ln\Gamma\left(\tfrac{m_j}{2}\right) - m_j \ln\Gamma\left(\tfrac{1}{2}\right)\right], \end{aligned} \tag{16}$$

kus Γ tähistab gammafunktsiooni, $\hat{n}_k = \sum_{i=1}^{n} \hat{z}_{ik}$ ning $\hat{u}_k^{jh} = \sum_{i=1}^{n} \hat{z}_{ik} x_i^{jh}$. Lisapaketis „Rmixmod“ on implementeeritud ainult asümptootiline ICL kriteerium (13) (Lebret jt, 2015, lk 10), kuid kuna kõigi hinnatud segumudelite klasterdused salvestatakse, saame iga mudeli korral arvutada ka $\mathrm{ICL}_{\mathcal{M}}$ väärtuse. Seega kasutame antud näites parima mudeli valimiseks nii asümptootilist kui ka täpset ICL kriteeriumi. Seejuures märgime, et erinevalt asümptootilisest kriteeriumist on täpse kriteeriumi korral parim mudel, mille $\mathrm{ICL}_{\mathcal{M}}$ väärtus on maksimaalne. Klasterdatavate vaatluste genereerimiseks ja klasteranalüüsi teostamiseks kirjutatud tarkvara R kood on toodud lisas 2.

Situatsioon 1

Esimeses situatsioonis kasutame $\delta = 0{,}4$. Siis valemiga (15) arvutatud vaatluse all oleva segujaotuse kahe komponendi parameetrite vektorid on kujul

$$\begin{aligned}\boldsymbol{\alpha}_1 &= \left(\frac{11}{15}, \frac{2}{15}, \frac{2}{15}; \quad \frac{11}{15}, \frac{2}{15}, \frac{2}{15}; \quad \frac{7}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}; \quad \frac{7}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}\right)', \\ \boldsymbol{\alpha}_2 &= \left(\frac{2}{15}, \frac{11}{15}, \frac{2}{15}; \quad \frac{2}{15}, \frac{11}{15}, \frac{2}{15}; \quad \frac{1}{10}, \frac{7}{10}, \frac{1}{10}, \frac{1}{10}; \quad \frac{1}{10}, \frac{7}{10}, \frac{1}{10}, \frac{1}{10}\right)'.\end{aligned}$$

Valimimahu $n = 1000$ korral on asümptootilise ICL kriteeriumi kohaselt kolm parimat segumudelit $[\varepsilon]$, $[\varepsilon_k^j]$ ja $[\varepsilon^j]$, kõik kahekomponendilised (ICL väärtused on vastavalt 8204, 8223,0 ja 8223,4). Kuigi üldjuhul võib mudel $[\varepsilon]$ tunduda liiga lihtne, on antud olukorras loomulik, et parim mudel on just $[\varepsilon]$, sest kolme ja nelja võimaliku väärtusega tunnuste korral on valemiga (15) arvutatud moodide tõenäosused samas suurusjärgus (hetkel vastavalt $11/15 \approx 0{,}73$ ja $7/10 = 0{,}7$). Parima mudeli $[\varepsilon]$ korral on klasterdamise veamäär 5% (ehk 50 vaatlust paigutati valesse klastrisse) ning $\hat{\gamma}_k^j \approx 0{,}707$ iga $j = 1, \ldots, 4$ ja $k = 1, 2$ korral. Täpne ICL kriteerium valib kahekomponendilised mudelid $[\varepsilon_k^j]$ ja $[\varepsilon^j]$. Mõlema mudeli korral on $\mathrm{ICL}_{\mathcal{M}} \approx -4059$, sest need mudelid klasterdavad vaatlused ühtemoodi. Parem mudel on seega $[\varepsilon^j]$, kuna selles mudelis on vähem parameetreid. Järelikult täpne ICL kriteerium valib genereeritud andmete tegelikule struktuurile vastava segumudeli. Mudeli $[\varepsilon^j]$ veamäär on 4,9% ning parameetrite $\boldsymbol{\alpha}_1$ ja $\boldsymbol{\alpha}_2$ hinnangud

on kujul

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_1 \approx (&0{,}72; 0{,}14; 0{,}14; \quad 0{,}712; 0{,}144; 0{,}144;\\ &0{,}7; 0{,}1; 0{,}1; 0{,}1; \quad 0{,}696; 0{,}101; 0{,}101; 0{,}101)',\\ \hat{\boldsymbol{\alpha}}_2 \approx (&0{,}14; 0{,}72; 0{,}14; \quad 0{,}144; 0{,}712; 0{,}144;\\ &0{,}1; 0{,}7; 0{,}1; 0{,}1; \quad 0{,}101; 0{,}696; 0{,}101; 0{,}101)'\end{aligned}$$

ehk tõenäosuste hinnangud on tegelikele väärtustele väga lähedased.

Suurema valimimahu $n = 3000$ korral on mõlema kriteeriumi kohaselt parimad kahekomponendilised mudelid $[\varepsilon^j]$, $[\varepsilon]$ ja $[\varepsilon_k^j]$. Nende mudelite asümptootilise ICL kriteeriumi väärtused on vastavalt 24031, 24032 ja 24081. Kõik kolm mudelit annavad tulemuseks sama vaatluste klasterduse, kus 149 vaatlust (4,97%) on valesti klasterdatud. Antud juhul on parim segumudel $[\varepsilon]$, kuna selle parameetrite arv on kõige väiksem. Selle mudeli korral on $\hat{\gamma}_k^j \approx 0{,}718$, kus $j = 1, \ldots, 4$ ja $k = 1, 2$.

Situatsioon 2

Teises situatsioonis olgu $\delta = 0{,}6$. Siis

$$\begin{aligned}\boldsymbol{\alpha}_1 &= (0{,}6; 0{,}2; 0{,}2; \quad 0{,}6; 0{,}2; 0{,}2; \quad 0{,}55; 0{,}15; 0{,}15; 0{,}15; \quad 0{,}55; 0{,}15; 0{,}15; 0{,}15)',\\ \boldsymbol{\alpha}_2 &= (0{,}2; 0{,}6; 0{,}2; \quad 0{,}2; 0{,}6; 0{,}2; \quad 0{,}15; 0{,}55; 0{,}15; 0{,}15; \quad 0{,}15; 0{,}55; 0{,}15; 0{,}15)'.\end{aligned}$$

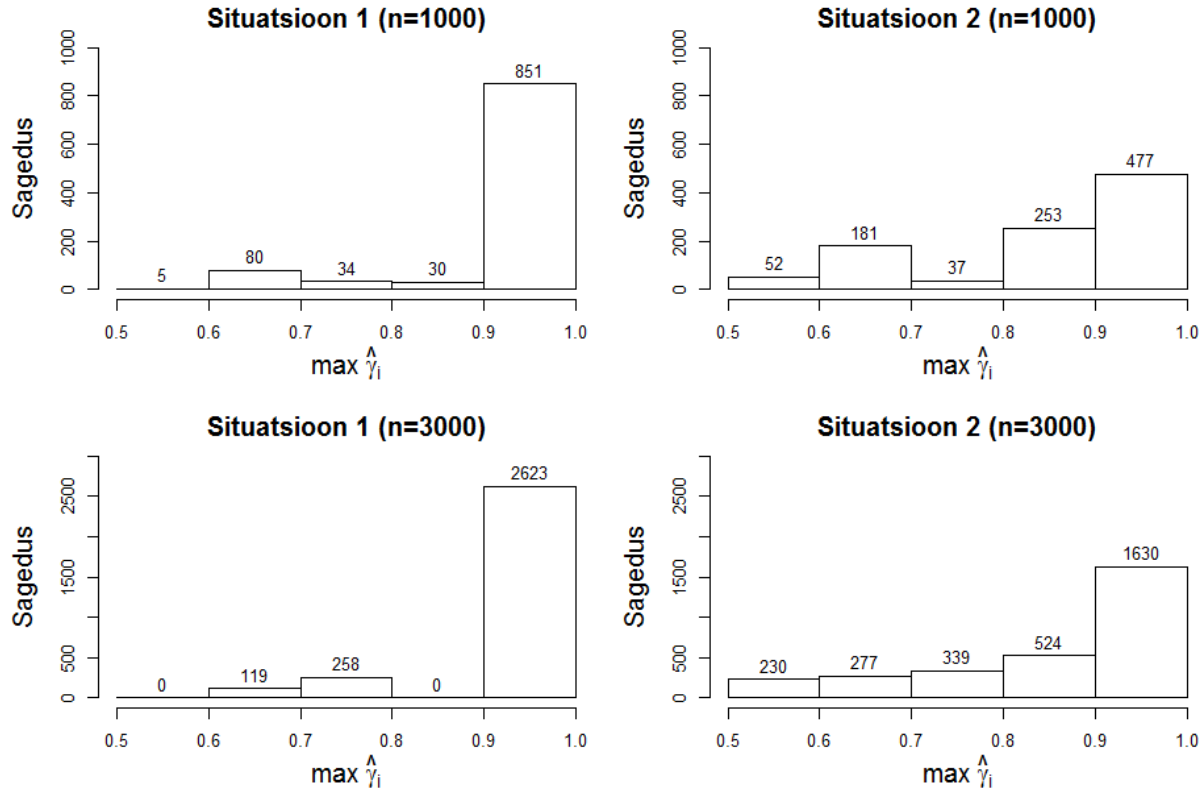
Kuna selles situatsioonis on klastrite kattuvus suurem, valivad mõlema genereeritud valimi korral mõlemad kriteeriumid ühe komponendiga segumudelid. Nende mudelite järel on valimimahu $n = 1000$ korral asümptootilise ICL kriteeriumi kohaselt parimad kahekomponendilised mudelid $[\varepsilon]$, $[\varepsilon^j]$ ja $[\varepsilon_k^j]$. Vastavad kriteeriumi väärtused on 9821, 9829 ja 9861 ning mudelite veamäärad on vastavalt 15%, 14,3% ja 14,3% (viimased kaks mudelit annavad sama vaatluste klasterduse). Kahekomponendiliste mudelite hulgast on täpse ICL kriteeriumi kohaselt parim andmete tegelikule struktuurile vastav mudel $[\varepsilon^j]$, mille korral näiteks tõenäosuste vektori $\boldsymbol{\alpha}_1$ hinnang on kujul

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_1 \approx (&0{,}603; 0{,}198; 0{,}198; \quad 0{,}599; 0{,}201; 0{,}201;\\ &0{,}537; 0{,}154; 0{,}154; 0{,}154; \quad 0{,}56; 0{,}147; 0{,}147; 0{,}147)'.\end{aligned}$$

Kui valimimahu $n = 3000$ korral ühekomponendilised segumudelid kõrvale jätta, on asümptootilise ICL kriteeriumi kohaselt parim mudel $[\varepsilon^j]$ ja täpse kriteeriumi kohaselt mudel $[\varepsilon]$, mõlemad kahekomponendilised. Nende kahe mudeli veamäärad on vastavalt 13,47% ja 13,43% ehk täpne ICL kriteerium valib natuke väiksema veamääraga mudeli (erinevus on kõigest üks vaatlus).

Genereeritud nelja valimi klastrite kattuvuse määra iseloomustab joonis 3 lk 23. Sellel

joonisel on kujutatud andmete tegelikule struktuurile vastavate kahekomponendiliste segumudelitega $[\varepsilon^j]$ saadud suuruste $\max \hat{\gamma}_i := \max\{\hat{\gamma}_1(\mathbf{x}_i;\boldsymbol{\theta}), \hat{\gamma}_2(\mathbf{x}_i;\boldsymbol{\theta})\}$, $i = 1, \ldots, n$, histogrammid. Suurus $\max \hat{\gamma}_i$ on tinglik tõenäosus, millele vastavasse klastrisse vaatlus $\mathbf{x}_i$ paigutatakse. Mida rohkem on vaatlusi, mille korral see suurus on lähedal ühele, seda selgemalt on vaatluse all olevad kaks klastrit üksteisest eraldatud. Jooniselt 3 on näha, et esimeses situatsioonis (kattuvusega $\delta = 0{,}4$) on mõlema valimimahu korral suurem osa suurustest $\max \hat{\gamma}_i$ lähedased ühele (suuremad kui 0,9). Teises situatsioonis ($\delta = 0{,}6$) on palju rohkem vaatluseid, mille korral tinglik tõenäosus $\max \hat{\gamma}_i$ on vahemikus $[0{,}5; 0{,}9]$. Seetõttu olid teises situatsioonis kahe komponendiga segumudelite klasterdamise veamäärad suuremad (13,5–15%) kui esimeses situatsioonis (5%).



Joonis 3. Suuruste $\max \hat{\gamma}_i := \max\{\hat{\gamma}_1(\mathbf{x}_i;\boldsymbol{\theta}), \hat{\gamma}_2(\mathbf{x}_i;\boldsymbol{\theta})\}$ $(i = 1, \ldots, n)$ histogrammid hinnatud kahekomponendiliste segumudelite $[\varepsilon^j]$ korral

Simulatsiooninäite tulemuste kohta võime kokkuvõtvalt öelda, et kvalitatiivsete andmete mudelipõhisel klasterdamisel tuleks parima segumudeli valimisel asümptootilise ICL kriteeriumi asemel eelistada täpset kriteeriumi. Täpne ICL kriteerium pole küll tarkvara R lisapaketis „Rmixmod“ realiseeritud, kuid selle arvutamine iga hinnatud segumudeli korral ei ole keeruline.

1.3 Segatüüpi andmete klasterdamine

Olgu vaatlusi $\mathbf{x}_1, \ldots, \mathbf{x}_n$ kirjeldavad l esimest tunnust kvantitatiivsed ning ülejäänud q mõõdetud tunnust kvalitatiivsed, $l + q = p$. Kvalitatiivsetel tunnustel olgu vastavalt $m_{l+1}, \ldots, m_p$ võimalikku väärtust. Kvantitatiivsete ja kvalitatiivsete tunnuste samaaegseks käsitlemiseks eeldatakse, et nende tunnuste rühmad on igas klastris sõltumatud. Mõlema rühma tunnuste modelleerimiseks kasutatakse tõenäosusjaotust. Segatüüpi andmete mudelipõhisel klasterdamisel eeldatakse, et iga vaatlus on sõltumatu realisatsioon segujaotusest

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \left[\pi_k f_k(\mathbf{x}^*; \boldsymbol{\theta}_k) \prod_{j=l+1}^{p} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x^{jh}} \right], \tag{17}$$

kus $\mathbf{x}^* \in \mathbb{R}^l$ on vektori $\mathbf{x}$ esimest l väärtust ehk mõõdetud kvantitatiivsete tunnuste väärtuseid sisaldav vektor ning ülejäänud tähistused on nagu segujaotustes (1) ja (14).

Segujaotuses (17) on samuti eeldatud, et kvalitatiivsed tunnused on lokaalselt sõltumatud ehk segujaotuse iga komponendi sees sõltumatud. Mõnikord tehakse see eeldus ka kvantitatiivsete tunnuste jaoks. Näiteks normaaljaotuste segu korral tähendab lokaalne sõltumatus, et komponentide kovariatsioonimaatriksid $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K$ on diagonaalsed. Sel juhul eeldatakse, et vaatluste kvantitatiivsete tunnuste väärtused pärinevad diagonaalsete kovariatsioonimaatriksitega normaaljaotuste segust ja kvalitatiivsete tunnuste väärtused pärinevad multinomiaalsete jaotuste segust. Segujaotuse (17) parameetrite hindamiseks kasutatakse EM-algoritmi ning segumudeli kuju ja komponentide arvu valimiseks kasutatakse asümptootilist ICL kriteeriumi.

1.4 Mudelipõhise lähenemise ja K-keskmiste meetodi võrdlus

Käesolevas peatükis esitame vaatlustevahelistel kaugustel põhineva K-keskmiste meetodi algoritmi ning kirjeldame kahte kriteeriumi optimaalse klastrite arvu leidmiseks kaugusi kasutavate klasterdamismeetodite korral. Peatükk lõpeb simulatsiooninäitega, kus võrdleme K-keskmiste meetodit ja mudelipõhist klasterdamist normaaljaotuste segu korral.

1.4.1 K-keskmiste meetod

Tänapäeval on üks kõige populaarsemaid klasteranalüüsi meetodeid kahtlemata K-keskmiste meetod, kuna see on kergesti rakendatav, kiire ja arvutuslikult efektiivne. Meetodi algoritm on järgmine (Izenman, 2008, lk 424):

1. Fikseeri klastrite arv K.

2. Paiguta vaatlused $\mathbf{x}_1, \ldots, \mathbf{x}_n$ klastritesse $C_1, \ldots, C_K$ ja arvuta klastrite keskpunktid $\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_K$.

3. Arvuta iga vaatluse ja vastava klastri keskpunkti vaheline eukleidilise kauguse ruut ning leia kauguste ruutude summa

$$L_K = \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)'(\mathbf{x}_i - \bar{\mathbf{x}}_k). \tag{18}$$

4. Paiguta iga vaatlus lähima keskpunktiga klastrisse. See garanteerib kaofunktsiooni L_K väärtuse vähenemise. Seejärel arvuta uued klastrite keskpunktid.

5. Korda samme 3 ja 4 kuni rohkem ümberpaigutusi ei toimu.

Kuna algoritm võib koonduda kaofunktsiooni (18) globaalse miinimumi asemel lokaalseks miinimumiks, tuleb meetodit rakendada mitu korda erinevaid vaatluste esialgseid konfiguratsioone kasutades. Saadud tulemustest parimaks valitakse klasterdus, mille kaofunktsiooni väärtus on minimaalne. Meetod on tundlik andmevigade ja erindite suhtes, kuna üksikud vigased või äärmiselt omapärased väärtused võivad klastrite keskpunkte märkimisväärselt nihutada ning seeläbi algoritmi tulemusi suurel määral mõjutada.

1.4.2 Optimaalse klastrite arvu leidmine

Antud alapeatükis vaatleme kahte kriteeriumi, mida kasutatakse optimaalse klastrite arvu leidmiseks kaugustel põhinevate klasteranalüüsi meetodite korral. Nendeks on keskmise silueti laiuse kriteerium ja gap-statistiku meetod. Tähistagu $\mathcal{C}_K = \{C_1, \ldots, C_K\}$ vaatluste $\mathbf{x}_1, \ldots, \mathbf{x}_n$ klasterdust, kus C_k on k-ndasse klastrisse paigutatud vaatluste indeksite hulk ning $n_k = |C_k|$ on klastrisse C_k kuuluvate vaatluste arv, $k = 1, \ldots, K$. Olgu d vaadeldav kaugusmõõt.

Keskmise silueti laiuse kriteerium

Keskmise silueti laiuse (*average silhouette width*) kriteerium põhineb klastritesisese homogeensuse ja klastritevahelise eraldatuse kompromissil (Kaufman ja Rousseeuw, 1990). Olgu vaatluse all klasterdus $\mathcal{C}_K$, $K > 1$. Tähistagu

$$a_i = \frac{1}{n_k - 1} \sum_{j \in C_k} d(\mathbf{x}_i, \mathbf{x}_j), \quad i = 1, \ldots, n,$$

klastrisse C_k kuuluva vaatluse $\mathbf{x}_i$ ja ülejäänud selle klastri vaatluste vahelist keskmist kaugust. Vaatluse $\mathbf{x}_i$ ja klastrisse C_l, $l \neq k$, paigutatud vaatluste vaheline keskmine kaugus

avaldub kujul

$$b_i^l = \frac{1}{n_l} \sum_{j \in C_l} d(\mathbf{x}_i, \mathbf{x}_j).$$

Seega $b_i = \min_{l \neq k} b_i^l$ on vaatluse $\mathbf{x}_i$ ja sellele järgmise lähima klastri vaheline keskmine kaugus. Vastavat klastrit kutsutakse vaatluse $\mathbf{x}_i$ naaberklastriks. Klasterduse $\mathcal{C}_K$ keskmine silueti laius arvutatakse valemiga

$$\bar{s}_K = \frac{1}{n} \sum_{i=1}^{n} s_i = \frac{1}{n} \sum_{i=1}^{n} \frac{b_i - a_i}{\max\{a_i, b_i\}}. \tag{19}$$

Silueti $s_i \in [-1, 1]$ väärtus kirjeldab, kui „hästi“ vaatlus $\mathbf{x}_i$ on klasterdatud (kui $n_k = 1$, siis $s_i = 0$). Näiteks $s_i \approx 1$ tähendab, et vaatlus $\mathbf{x}_i$ on hästi klasterdatud, sest klaster, kuhu see vaatlus kuulub, on homogeenne ($a_i \approx 0$). Teisisõnu, $\mathbf{x}_i$ on oma praeguse klastri vaatlustele keskmiselt palju lähemal kui naaberklastri vaatlustele: $a_i \ll b_i$. Kui $s_i \approx 0$, siis $a_i \approx b_i$ ehk vaatlus $\mathbf{x}_i$ paikneb praeguse klastri ja naaberklastri vahel. Kui $s_i \approx -1$ ehk $b_i \ll a_i$, siis $\mathbf{x}_i$ on halvasti klasterdatud. Järelikult, mida lähemal keskmine silueti laius (19) on ühele, seda optimaalsem on klastrite arv K. Seega, kui on moodustatud klasterdused $\mathcal{C}_2, \ldots, \mathcal{C}_{K_{\max}}$, kus $K_{\max} < n$, ja leitud vastavad keskmise silueti laiused $\bar{s}_2, \ldots, \bar{s}_{K_{\max}}$, siis optimaalne klastri arv on $\hat{K} = \arg\max_K \bar{s}_K$.

Gap-statistiku meetod

Käesolev alapunkt põhineb artiklil Tibshirani, Walther ja Hastie (2001) ning raamatul Hastie, Tibshirani ja Friedman (2009). Olgu

$$D_k = \sum_{i,j \in C_k} d(\mathbf{x}_i, \mathbf{x}_j), \quad k = 1, \ldots, K,$$

klastrisse C_k kuuluvate vaatluste paariviisiliste kauguste summa ja

$$W_K = \sum_{k=1}^{K} \frac{1}{2n_k} D_k$$

klastritesisest varieeruvust iseloomustav suurus (edaspidi hajuvus). Näiteks eukleidilise kauguse ruudu korral on see hajuvus kujul $W_K = \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)'(\mathbf{x}_i - \bar{\mathbf{x}}_k)$, sest siis $D_k = 2n_k \sum_{i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)'(\mathbf{x}_i - \bar{\mathbf{x}}_k)$. Klastrite arvu suurenedes hajuvuse W_K väärtus väheneb monotoonselt, aga mingist arvust alates see vähenemine aeglustub märgatavalt. Olgu andmetes tegelikult K^* üksteisest selgesti eraldatud gruppi. Kui $K < K^*$, siis mõned leitud klastrid koosnevad mitmest grupist pärinevatest vaatlustest. Sel juhul $W_{K+1} \ll W_K$, sest klastrite arvu suurenedes saab rohkem tegelikest gruppidest endale oma klastri. Kui $K > K^*$, siis vähemalt üks tegelik grupp on kaheks klastriks jagatud ja klastrite arvu edasine suurendamine hajuvuse W_K väärtust palju ei vähenda. Teisisõnu, ühe tegeliku

grupi osadeks jagamine vähendab suurust W_K vähem kui kahe üksteisest selgesti eraldatud grupi eraldamine: $\{W_K - W_{K+1} \,|\, K < K^*\} \gg \{W_K - W_{K+1} \,|\, K \geq K^*\}$.

Gap-statistiku meetod seisneb suuruse $\ln W_K$ võrdlemisel selle keskväärtusega sobivalt valitud (andmestiku kuju poolt määratud) ühtlase jaotuse U korral ehk olukorras, kus tegelikult klastreid ei ole. Gap-statistik defineeritakse valemiga

$$G(K) = E_U(\ln W_K) - \ln W_K. \tag{20}$$

Keskväärtuse $E_U(\ln W_K)$ hinnangu arvutamiseks genereeritakse B valimit ühtlasest jaotusest U, arvutatakse iga valimi korral suurus $\ln W_K$ ning leitakse saadud väärtuste $\ln W_K^{(1)}, \ldots, \ln W_K^{(B)}$ keskmine. Gap-statistiku meetodi korral on optimaalne klastrite arv kõige väiksem K väärtus, mille korral $G(K) \geq G(K+1) - s_{K+1}$, kus

$$s_K = \mathrm{sd}(K)\sqrt{1 + 1/B}$$

ning $\mathrm{sd}(K)$ on väärtuste $\ln W_K^{(b)}$, $b = 1, \ldots, B$, standardhälve. Kokkuvõttes on meetodi algoritm järgmine:

1. Moodusta vaatluste $\mathbf{x}_1, \ldots, \mathbf{x}_n$ klasterdused $\mathcal{C}_1, \ldots, \mathcal{C}_{K_{\max}}$ ja arvuta hajuvused $W_1, \ldots, W_{K_{\max}}$.

2. Genereeri B valimit sobivalt valitud ühtlasest jaotusest. Klasterda iga valim ja leia gap-statistiku (20) hinnang

$$\hat{G}(K) = \frac{1}{B}\sum_{b=1}^{B} \ln W_K^{(b)} - \ln W_K, \quad K = 1, \ldots, K_{\max}.$$

3. Arvuta $\bar{l}_K = \frac{1}{B}\sum_{b=1}^{B} \ln W_K^{(b)}$ ning $s_K = \mathrm{sd}(K)\sqrt{1 + 1/B}$, kus

$$\mathrm{sd}(K) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\ln W_K^{(b)} - \bar{l}_K\right)^2}, \quad K = 1, \ldots, K_{\max}.$$

4. Vali $\hat{K}^* = \arg\min\{K \,|\, \hat{G}(K) \geq \hat{G}(K+1) - s_{K+1}\}$.

Tarkvaras R on gap-statistik realiseeritud paketi „cluster“ funktsioonis `clusGap()`.

1.4.3 *K*-keskmiste meetod kui mudelipõhise klasterdamise erijuht

Alapeatükis 1.4.1 kirjeldatud K-keskmiste meetod leiab võimalikult homogeensed klastrid, kus iga vaatlus on paigutatud klastrisse, mille keskpunktile see on eukleidilise kauguse ruudu mõttes kõige lähemal. See tähendab, et K-keskmiste meetod otsib

sfäärilisi klastreid ja ei tööta hästi, kui vaatlused moodustavad näiteks piklikud grupid ehk kui gruppidesisene hajuvus on mõnes suunas kordades suurem kui ülejäänud suundades. Selgub, et K-keskmiste meetod ja mudelipõhine klasteranalüüs normaaljaotuste segu korral on samaväärsed, kui segujaotuse komponentide kaalud on võrdsed ja komponentidel on sama sfääriline kovariatsioonimaatriks (Celeux ja Govaert, 1992).

Olgu vaatluse all segujaotus (1), mille komponentide tihedused f_k on antud valemiga (4), kus $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}_p$, $\sigma^2 > 0$, ja $\pi_k = 1/K$ iga $k = 1, \ldots, K$ korral. Lisaks olgu $C_1, \ldots, C_K$ tundmatute indikaatorvektorite $\mathbf{z}_1, \ldots, \mathbf{z}_n$ poolt määratud vaatluste $\mathbf{x}_1, \ldots, \mathbf{x}_n$ klasterdus ning $n_k = |C_k|$, $k = 1, \ldots, K$. Siis vaatluste $\{\mathbf{x}, \mathbf{z}\} = \{(\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n)\}$ logaritmiline tõepärafunktsioon (7) on kujul

$$\begin{aligned}
\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) &= \sum_{k=1}^{K} \sum_{i \in C_k} \ln \left[\frac{1}{K} \frac{1}{(2\pi)^{p/2} |\sigma^2 \boldsymbol{I}_p|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)' (\sigma^2 \boldsymbol{I}_p)^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \right] \\
&= \sum_{k=1}^{K} n_k \ln \left[K^{-1} (2\pi)^{-p/2} (\sigma^2)^{-p/2} \right] - \frac{1}{2\sigma^2} \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \boldsymbol{\mu}_k)' (\mathbf{x}_i - \boldsymbol{\mu}_k) \\
&= A - \frac{np}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \boldsymbol{\mu}_k)' (\mathbf{x}_i - \boldsymbol{\mu}_k),
\end{aligned}$$

kus $A = -n \ln(K) - \frac{np}{2} \ln(2\pi)$ on konstant. Lihtne on veenduda, et keskväärtusvektori $\boldsymbol{\mu}_k$ suurima tõepära hinnang on klastri C_k keskpunkt $\bar{\mathbf{x}}_k$, $k = 1, \ldots, K$, ja ühise dispersiooni σ^2 suurima tõepära hinnang on

$$\hat{\sigma}^2 = \frac{1}{np} \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)' (\mathbf{x}_i - \bar{\mathbf{x}}_k).$$

Seega avaldub maksimeeritud logaritmiline tõepära klasterduse $C_1, \ldots, C_K$ korral kujul

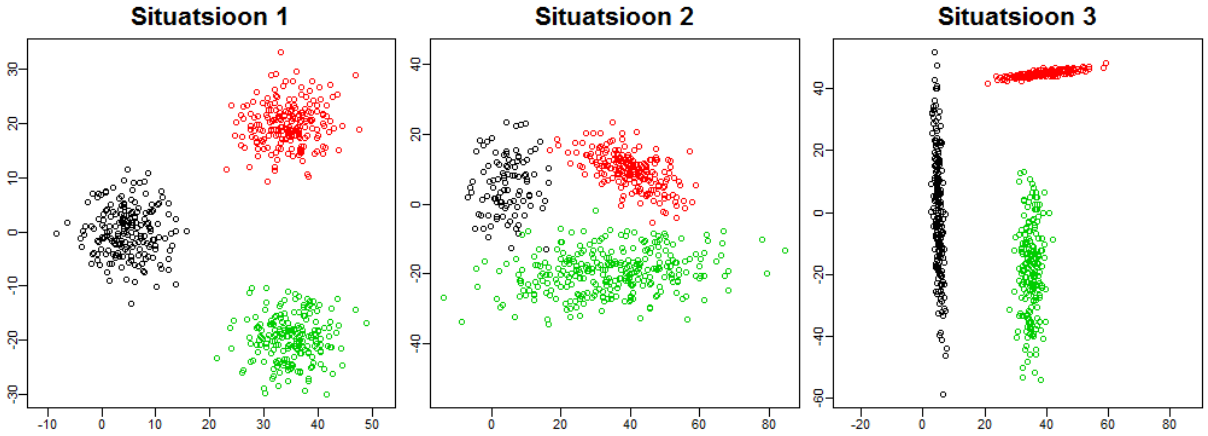
$$\ln p(\mathbf{x}, \mathbf{z}; \hat{\boldsymbol{\theta}}) = A - \frac{np}{2} \ln \left[\frac{1}{np} \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)' (\mathbf{x}_i - \bar{\mathbf{x}}_k) \right] - \frac{np}{2}$$

ehk funktsioon $\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ on maksimaalne, kui summa $\sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)' (\mathbf{x}_i - \bar{\mathbf{x}}_k)$ on minimaalne. See tähendabki, et logaritmilise tõepära $\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ maksimeerimine on samaväärne K-keskmiste meetodi kaofunktsiooni (18) minimeerimisega.

Näide 3. Kvantitatiivsete andmete klasterdamine erineva kujuga klastrite korral

Järgnevalt illustreerime K-keskmiste meetodi ja mudelipõhise klasteranalüüsi käitumist erineva kujuga klastrite korral. Selleks vaatleme kolme erinevat situatsiooni, iga kord genereerime 600 vaatlust kolme komponendiga kahemõõtmeliste normaaljaotuste segust. Esimeses situatsioonis on kõik kolm kovariatsioonimaatriksit sfäärilised (diagonaalsed ja

ühise dispersiooniga) ning võrsed, keskväärtusvektorid on valitud nii, et grupid omavahel ei kattuks. Kirjeldatud segujaotusest genereeritud vaatlused (igast komponendist 200) moodustavad kolm üksteisest eraldatud võrdse suurusega sfäärilist gruppi, seetõttu peaksid mõlemad meetodid tegelikud grupid üles leidma. Teisel juhul on normaaljaotuste parameetrid valitud nii, et vaatluste grupid oleksid selgesti eristatavad, aga gruppidevaheline kaugus oleks võimalikult väike. Ükski kovariatsioonimaatriks ei ole diagonaalne, kusjuures ühes grupis on hajuvus ühes suunas mitu korda suurem kui teises suunas. Lisaks on grupid erineva suurusega. Sel juhul peaks mudelipõhine lähenemine sobivam olema, kuid kuna ainult üks grupp on piklik, võiks ka K-keskmiste meetod tõele lähedase tulemuse anda. Kolmandas situatsioonis on grupid uuesti võrdse suurusega. Kõigi kolme kovariatsioonimaatriksi korral on ühe tunnuse dispersioon mitukümmend või isegi mitusada korda suurem kui teise tunnuse oma ehk genereeritud vaatlused moodustavad väga välja venitatud grupid. Keskväärtusvektorid on valitud nii, et grupid oleksid üksteisest selgesti eraldatud. Viimases situatsioonis peaks tõele lähedase tulemuse andma ainult mudelipõhine klasteranalüüs, kuna K-keskmiste meetod otsib sfäärilisi klastreid. Kirjeldatud kolme situatsiooni kohaselt genereeritud vaatlused on toodud joonisel 4.



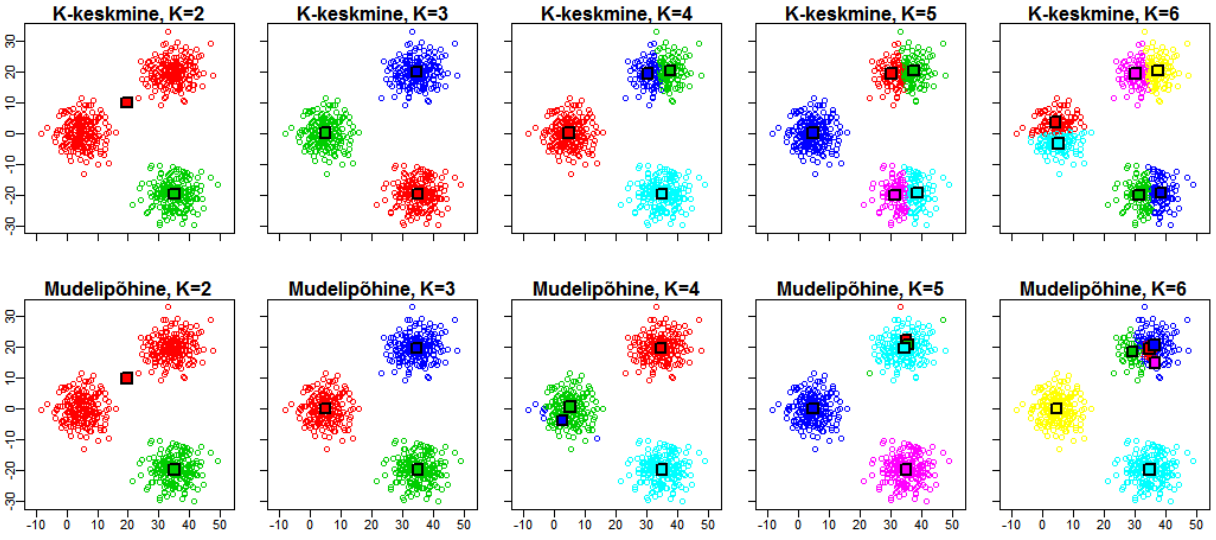
Joonis 4. Kolme situatsiooni kohaselt genereeritud vaatlused

Mõlema vaatluse all oleva meetodi korral teostame klasteranalüüsi klastrite arvudega $K = 2, \ldots, 6$. Mudelipõhise klasteranalüüsi korral rakendame EM-algoritmi 20 korda erinevate alglähenditega (maksimaalne iteratsioonide arv on 1000 ja logaritmilise tõepära suhtelise muutuse koondumiskriteeriumis (6) on $\varepsilon = 0{,}001$), K-keskmiste meetodi algoritmi rakendame 50 korda (maksimaalne iteratsioonide arv on 100). Optimaalse klastrite arvu valimiseks K-keskmiste meetodi korral kasutame keskmise silueti laiuse kriteeriumi ja gap-statistiku meetodit ($K_{\max} = 6$ ja $B = 600$), mudelipõhise lähenemise korral asümptootilist ICL kriteeriumi (13). Kuigi mudelipõhine klasteranalüüs ei ole keskmise silueti laiuse kriteeriumiga kooskõlas, sest viimane põhineb vaatlustevahelistel

kaugustel, leiame eukleidilise kauguse ruutu kasutades võrdluseks ka selle kriteeriumi väärtused mudelipõhise klasteranalüüsi tulemuste jaoks. Joonisel 4 lk 29 kujutatud vaatluste genereerimiseks ja klasteranalüüsi teostamiseks kirjutatud tarkvara R kood on toodud lisas 3.

Situatsioon 1

Esimeses situatsioonis genereerisime 200 vaatlust kolmest normaaljaotusest keskväärtusvektoritega $\boldsymbol{\mu}_1 = (5, 0)'$, $\boldsymbol{\mu}_2 = (35, 20)'$ ja $\boldsymbol{\mu}_3 = (35, -20)'$ ning sfääriliste kovariatsioonimaatriksitega $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = 20\boldsymbol{I}_2$. Mõlema meetodi tulemused kõigi proovitud klastrite arvude korral on kujutatud joonisel 5, kus värviliste ruutudega on K-keskmiste meetodi korral tähistatud klastrite keskpunktid ning mudelipõhise lähenemise korral keskväärtusvektorite $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_3$ hinnangud. Sellelt jooniselt on näha, et kahe ja kolme klastri korral on meetodite tulemused samad, kusjuures kolme klastri korral on ootuspäraselt saadud klastrid ja tegelikud grupid üksüheses vastavuses (võrdle joonisega 4 lk 29). Suuremate klastrite arvude korral on näha, kuidas K-keskmiste meetod poolitab tegelikud grupid ligikaudu võrdse suurusega klastriteks. Mudelipõhise lähenemise korral aga otsitakse normaaljaotusega komponente ja seetõttu võib väike hulk vaatlusi määrata omaette klastri, mis juhtus nelja, viie ning kuue klastri korral. Kui mudelipõhise klasteranalüüsi korral on klastrite arv liiga suur, võivad tekkida väga väikeste dispersioonidega ehk kunstlikud komponendid. Näiteks kuue klastri korral on viiendas (lillas) klastris $\hat{\boldsymbol{\Sigma}}_5 \approx \text{diag}(0{,}02; 0{,}17)$. Selles klastris on kaheksa vaatlust.



Joonis 5. Klasteranalüüsi tulemused situatsioonis 1

Antud situatsioonis on mõlema meetodi korral kolm klastrit optimaalne (vt tabel 1 lk 31). Gap-statistiku kohaselt on samuti kolm klastrit optimaalne.

Tabel 1. Keskmise silueti laiuse $\bar{s}_K$ ja ICL väärtused situatsioonis 1

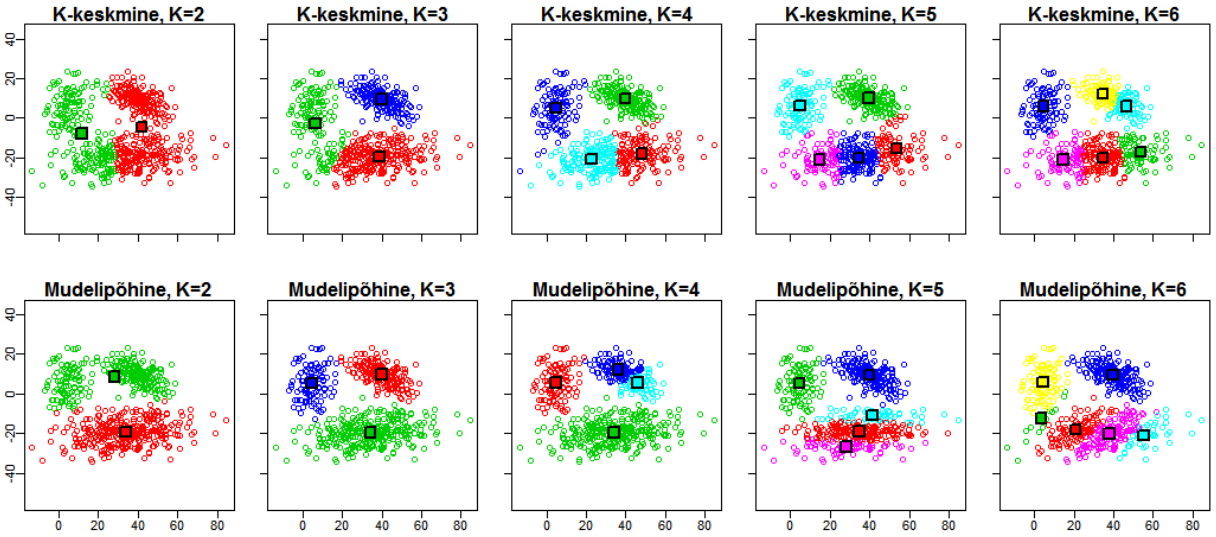
Meetod	Kriteerium	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
K-keskmine	$\bar{s}_K$	0,656	**0,937**	0,779	0,620	0,472
Mudelipõhine	$\bar{s}_K$	0,656	**0,937**	0,784	0,811	0,613
	ICL	8883	**8266**	8445	8482	8591

Situatsioon 2

Teises situatsioonis kasutasime parameetreid $\boldsymbol{\mu}_1 = (5,5)'$, $\boldsymbol{\mu}_2 = (40,10)'$, $\boldsymbol{\mu}_3 = (35,-20)'$,

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 35 & 5 \\ 5 & 75 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 65 & -25 \\ -25 & 30 \end{pmatrix}, \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 250 & 15 \\ 15 & 40 \end{pmatrix}.$$

Esimesest normaaljaotusest genereerisime 100, teisest 200 ning kolmandast 300 vaatlust. Genereeritud vaatluste kõik klasterdused on toodud joonisel 6. Antud situatsioonis on kõigi klastrite arvude korral K-keskmiste meetodi ja mudelipõhise lähenemise tulemused erinevad. Põhjuseks on gruppidevaheline väike kaugus ning kolmanda grupi vaatluste esimese tunnuse suur dispersioon. Seetõttu on näiteks kolme klastri korral K-keskmiste meetod paigutanud osa kolmanda grupi vaatluseid esimese grupi vaatlustega samasse klastrisse. Mudelipõhise klasterdamise korral on ainult väike hulk vaatlusi kolme klastri korral valesti klasterdatud (võrdle joonisega 4 lk 29). Jooniselt 6 on samuti selgesti näha, kuidas K-keskmiste meetod otsib sfäärilisi klastreid, aga mudelipõhise lähenemise korral sellist piirangut klastrite kujule ei ole.



Joonis 6. Klasteranalüüsi tulemused situatsioonis 2

Tabelist 2 lk 32 selgub, et K-keskmiste meetodi korral on keskmine silueti laius suurim nelja klastri korral, kuid kolme ja viie klastri korral ei ole kriteeriumi väärtused palju

väiksemad. Gap-statistiku kohaselt on hoopis üks klaster optimaalne. Asümptootilise ICL kriteeriumi kohaselt on mudelipõhise lähenemise korral parim kolmekomponendiline segumudel. Järelikult töötas selles situatsioonis tõepoolest mudelipõhine klasteranalüüs paremini kui K-keskmiste meetod.

Tabel 2. Keskmise silueti laiuse $\bar{s}_K$ ja ICL väärtused situatsioonis 2

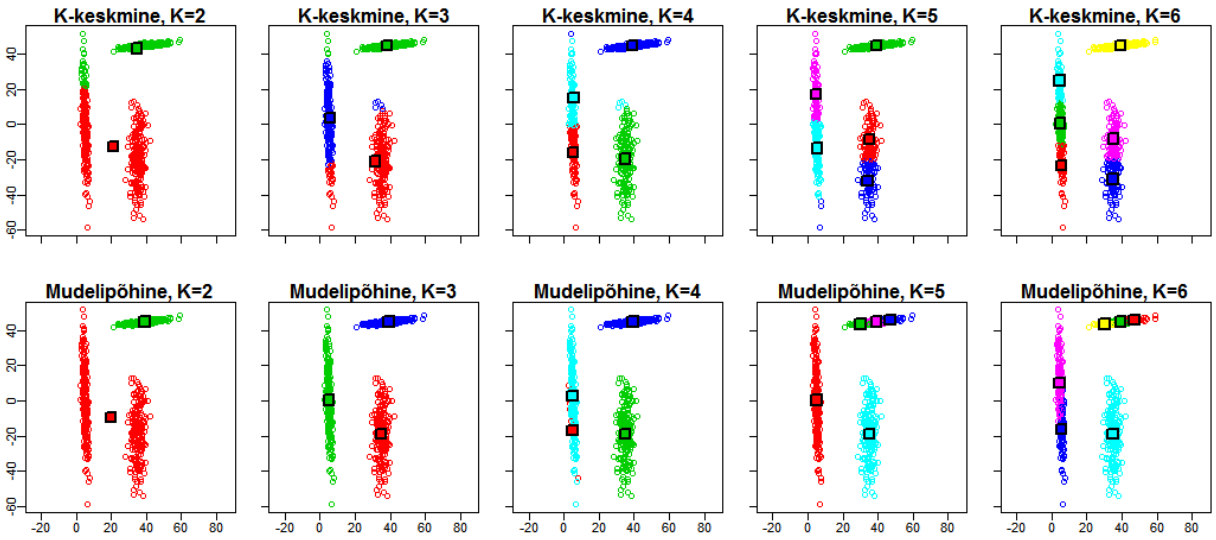
Meetod	Kriteerium	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
K-keskmine	$\bar{s}_K$	0,550	0,657	**0,690**	0,657	0,596
Mudelipõhine	$\bar{s}_K$	0,544	**0,619**	0,457	0,400	0,534
	ICL	9876	**9604**	9773	9801	9884

Situatsioon 3

Kolmanda situatsiooni jaoks võtsime $\boldsymbol{\mu}_1 = (5, 0)'$, $\boldsymbol{\mu}_2 = (40, 45)'$, $\boldsymbol{\mu}_3 = (35, -20)'$ ja

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & -10 \\ -10 & 400 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 45 & 5 \\ 5 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 5 & 1 \\ 1 & 235 \end{pmatrix}.$$

Igast jaotusest genereerisime 200 vaatlust. Kõigi teostatud klasteranalüüside tulemused on kujutatud joonisel 7. Sellelt jooniselt on näha, et parim kolme komponendiga segumudel klasterdab kõik vaatlused õigesti. Kuna esimene ja kolmas grupp on väga pikaks venitatud ja gruppidevaheline kaugus pole piisavalt suur, paigutab K-keskmiste meetod kolme klastri korral märkimisväärse hulga vaatlusi valedesse klastritesse.



Joonis 7. Klasteranalüüsi tulemused situatsioonis 3

Tabelist 3 lk 33 selgub, et K-keskmiste meetodi korral on keskmine silueti laius kõigi vaadeldud klastrite arvude korral samas suurusjärgus. On selge, et K-keskmiste meetod selles situatsioonis ei tööta. Gap-statistiku kohaselt on genereeritud andmetes viis

klastrit. Mudelipõhise klasteranalüüsi korral on keskmine silueti laius suurim kahe ja kolme klastri korral ning ICL väärtus on minimaalne kolme klastri korral ehk parim on kolmekomponendiline segumudel. Seega töötas selles situatsioonis ootuspäraselt samuti mudelipõhine lähenemine paremini.

Tabel 3. Keskmise silueti laiuse $\bar{s}_K$ ja ICL väärtused situatsioonis 3

Meetod	Kriteerium	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
K-keskmine	$\bar{s}_K$	0,760	0,746	0,763	0,762	0,760
Mudelipõhine	$\bar{s}_K$	0,729	**0,733**	0,514	0,575	0,552
	ICL	9239	**7867**	8003	8098	8190

Antud näites klasterdasime erinevatest normaaljaotustest genereeritud vaatlusi. Seetõttu ei olnud üllatav, et K-keskmiste meetod töötas mudelipõhise lähenemisega võrreldes sama hästi ainult esimeses situatsioonis, kus kõik kolm kovariatsioonimaatriksit olid sfäärilised ja grupid olid üksteisest selgesti eraldatud. Samuti polnud üllatuseks K-keskmiste meetodi tulemuste suuremad keskmise silueti laiuse väärtused teises ja kolmandas situatsioonis, sest see kriteerium on mõeldud optimaalse klastrite arvu leidmiseks kaugustel põhinevate meetodite korral.

1.5 Tarkvara R lisapakett „Rmixmod“

Tarkvaras R on mudelipõhist klasteranalüüsi lihtne rakendada lisapaketiga „Rmixmod“. Antud töös kasutame selle paketi versiooni 2.1.2. Järgnev ülevaade paketi „Rmixmod“ funktsioonidest põhineb artiklil Lebret jt (2015) ja paketi dokumentatsioonil Langrognet jt (2018).

Mudelipõhist klasteranalüüsi teostab funktsioon `mixmodCluster()`, mille olulisemad argumendid on `data`, `nbCluster`, `dataType`, `models`, `strategy` ning `criterion`. Neist kohustuslikud on ainult klasterdatavatel vaatlustel mõõdetud tunnuste väärtusi sisaldav andmestik `data`, kus on n rida ja p veergu, ning hinnatavate klastrite arvude vektor `nbCluster`. Argumendiga `dataType` saab määrata, kas vaatlusi iseloomustavad tunnused on kvantitatiivsed (`"quantitative"`), kvalitatiivsed (`"qualitative"`) või segatüüpi (`"composite"`). Vaikimisi on selle argumendi väärtus `NULL` ja tunnuste tüüp määratakse andmestiku `data` põhjal.

Argumendiga `models` määratakse hinnatavad segumudelid. Kvantitatiivsete tunnuste korral eeldatakse normaaljaotuste segu. Sel juhul saab valida erinevate kitsendustega 28 segumudeli hulgast. Kitsenduste seadmisel lähtutakse kovariatsioonimaatriksi esitusest

(5). Lubades osadel selle dekompositsiooni teguritel segujaotuse komponentide vahel varieeruda ja fikseerides ülejäänud, saadakse kaheksa erinevat mudelit. Neid tähistatakse järgmiselt: $[\lambda_k D_k A_k D_k']$, $[\lambda D_k A_k D_k']$, $[\lambda_k D_k A D_k']$, $[\lambda D_k A D_k']$, $[\lambda_k D A_k D']$, $[\lambda D A_k D']$, $[\lambda_k D A D']$ ja $[\lambda D A D']$. Näiteks mudel $[\lambda D_k A D_k']$ tähendab, et saadavad klastrid on sama suuruse ja kujuga, aga nende suund võib olla erinev. Diagonaalsete kovariatsioonimaatriksite korral on võrdus (5) kujul $\mathbf{\Sigma}_k = \lambda_k B_k$, kus B_k on konstandiga $\lambda_k = |\mathbf{\Sigma}_k|^{1/p}$ normaliseeritud maatriksi $\mathbf{\Sigma}_k$ omaväärtuste diagonaalmaatriks. Sellest võrdusest saadakse järgmised neli mudelit: $[\lambda_k B_k]$, $[\lambda B_k]$, $[\lambda_k B]$, $[\lambda B]$. Lõpuks, eeldades sfäärilisi kovariatsioonimaatrikseid ($A_k = \boldsymbol{I}_p \ \forall k$), lisanduvad veel mudelid $[\lambda_k \boldsymbol{I}_p]$, $[\lambda \boldsymbol{I}_p]$. Kuna pakett „Rmixmod“ võimaldab ka segujaotuse komponentide kaalud võrdseks kitsendada ($\pi_k = 1/K \ \forall k$), ongi kvantitatiivsete tunnuste korral vaatluse all 28 erinevat mudelit.

Kui kõik klasterdamiseks kasutatavad tunnused on kvalitatiivsed, eeldab funktsioon `mixmodCluster()` multinomiaalsete jaotuste segu (14). Paketis on kasutusel peatüki 1.2 kitsenduste seadmise alapunktis ära toodud parametrisatsioon ehk realiseeritud on segumudelid $[\varepsilon_k^{jh}]$, $[\varepsilon_k^j]$, $[\varepsilon_k]$, $[\varepsilon^j]$, $[\varepsilon]$. Arvestades jälle võimalusega, et komponentide kaalud saab võtta võrdseks, on kvalitatiivsete tunnuste korral võimalik hinnata kümme erinevat mudelit. Segatüüpi andmete klasterdamisel eeldatakse, et kõik tunnused on iga klastri sees sõltumatud. Siis on 28 kvantitatiivsete tunnuste segumudeli hulgast vaatluse all kaheksa mudelit, kus kovariatsioonimaatriksid on diagonaalsed, kuid mitte sfäärilised. Nende kaheksa ja viie kvalitatiivsete tunnuste mudelite kõikvõimalikud kombinatsioonid annavad tulemuseks 40 erinevat segumudelit. Kõigi võimalike segumudelite hindamiseks tuleb sõltuvalt kasutatavate tunnuste tüübist argumendi `models` väärtuseks panna `mixmodGaussianModel()`, `mixmodMultinomialModel()` või `mixmodCompositeModel()`.

Funktsiooni `mixmodCluster()` järgmine oluline argument on `strategy`, mille abil saab muuta segujaotuste parameetrite hindamiseks kasutatavat strateegiat. Selleks tuleb kasutada funktsiooni `mixmodStrategy()`, mille olulisemad argumendid on parameetrite hindamiseks kasutatav algoritm `algo` (vaikimisi `"EM"`), selle algoritmi rakendamise kordade arv `nbTry` (vaikimisi 1) ja algoritmi alglähendite leidmise meetod `initMethod`. Vaikimisi leitakse parameetrite alglähendid niinimetatud lühikese EM-algoritmi (`initMethod="smallEM"`) rakendamise kaudu. Selle meetodi korral valitakse juhuslikult parameetrite alglähendid ja teostatakse väike arv EM-algoritmi iteratsioone (vaikimisi `nbIterationInInit=5`). Antud protsessi korratakse kuni teostatud iteratsioonide arv on argumendi `nbTryInInit` väärtusest (vaikimisi 50) väiksem. Lõplikeks alglähenditeks valitakse saadud parameetrite hinnangute seast need, mille korral logaritmilise tõepärafunktsiooni väärtus on kõige suurem. Parameetrite hindamiseks kasutatava algoritmi `algo` koondumistingimusi määravateks funktsiooni `mixmodStrategy()` argumentideks

on maksimaalne iteratsioonide arv `nbIterationInAlgo` (vaikimisi 200) ning konstandi ε väärtus `epsilonInAlgo` logaritmilise tõepärafunktsiooni suhtelise muutuse kriteeriumis (6) (vaikimisi 0,001).

Funktsiooni `mixmodCluster()` viimane oluline argument `criterion` määrab kriteeriumi, mida kasutatakse parima segumudeli valimiseks. Selle argumendi võimalikud väärtused on `"ICL"` (asümptootiline ICL kriteerium (13)), `"BIC"` (Bayesi kriteerium (9)) ja `"NEC"` (niinimetatud normaliseeritud entroopia kriteerium, mida antud töös ei käsitleta). Vaikimisi kasutab funktsioon Bayesi kriteeriumi. Funktsiooni `mixmodCluster()` rakendamise tulemuseks on objekt, mis sisaldab kõigi hinnatud segumudelite tulemuste (näiteks parameetrite hinnangud, maksimeeritud logaritmiline tõepära, vaatluste klasterdus) loendit `results`, kus mudelid on sorteeritud kriteeriumi `criterion` väärtuste järgi. Selle loendi esimese ehk parima mudeli tulemused salvestatakse eraldi ka objekti `bestResult`. Tulemuste reprodutseerimiseks saab funktsioonis `mixmodCluster()` argumendiga `seed` fikseerida kasutatava juhusliku seemne.

Funktsiooni `mixmodCluster()` suurimaks puuduseks on, et segatüüpi tunnuste korral eeldatakse, et kvantitatiivsed tunnused on samuti lokaalselt sõltumatud (kovariatsioonimaatriksid on diagonaalsed). Teiseks oluliseks puuduseks on, et parameetrite hindamisel teostatud iteratsioonide arvu ei väljastata, mis raskendab hindamisstrateegiaga seotud parameetrite valimist.

2 Geenivaramu andmete klasteranalüüs

Käesolevas peatükis rakendame mudelipõhist klasteranalüüsi segatüüpi andmete korral. Analüüsitavad andmed pärinevad Tartu Ülikooli Eesti Geenivaramust. Vaatluse all olevat andmestikku kasutati Geenivaramu ja Soome molekulaarse meditsiini instituudi teadlaste ühisuuringus, kus püüti kindlaks teha suremusriski mõjutavaid biomarkereid. Järgnev ülevaade sellest uuringust põhineb artiklil Fischer jt (2014) ning pressiteatel Allik (2014).

Biomarkerid on veres, teistes kehavedelikes ja kudedes leiduvad biomolekulid, mis võivad märku anda ebaharilikest protsessidest või haigustest organismis. Näiteks üldkolesterooli hulk veres on üks peamine südame-veresoonkonnahaiguste tekkimise riski kirjeldav biomarker. Biomarkereid kasutatakse põhiliselt konkreetsete haiguste väljakujunemise riski hindamiseks. Inimese üldist tervislikku seisundit hästi iseloomustavaid ning suremusriski prognoosida aitavaid biomarkereid seni avastatud ei ole. Mainitud uuringu eesmärk oli tuvastada suremusega seotud biomarkereid, mis aitaksid omavahel erinevaid haiguseid siduda ja paremini hinnata inimeste suremusriski. Vaatluse all oli 106 markerit, tegemist oli erinevate ainete, nagu näiteks valgud ja aminohapped, kontsentratsioonidega veres. Uuringu TÜ Eesti Geenivaramu kohordi moodustasid ajavahemikul 9. oktoober 2002 kuni 16. veebruar 2011 geenidoonoriks hakanud 50715 indiviidi. Nendest 9842 kaasati juhuslikult uuringusse ja iga indiviidi vereplasmast määrati 106 huvipakkuva biomarkeri sisaldused. Andmete modelleerimisel tuvastati neli suremusega seotud biomarkerit, mis tulevikus võivad olla olulised näitajad inimese üldise tervisliku seisundi hindamisel ja parema ravi määramisel. Nendeks markeriteks olid albumiin, α-1 glükoproteiin, tsitraat ja VLDL-partikli diameeter. Näiteks nii albumiini madal kui ka glükoproteiini kõrge tase võivad viidata erinevatele põletikulistele protsessidele organismis, probleemidele neerude või maksa töös. Tuvastatud nelja markeri analüüs 7503 indiviidil Soome FINRISK 1997 uuringus kinnitas, et tegu polnud juhuleiuga.

2.1 Andmestiku ülevaade

Magistritöös vaatluse all olev andmestik on osavalim kirjeldatud uuringus analüüsitud Geenivaramu kohordi valimist, mis sisaldab 9787 geenidoonori andmeid. Geenidoonorite klasterdamiseks kasutatavate tunnuste hulgas on seitse biomarkerit (HDL-kolesterool ehk niinimetatud hea kolesterool, triglütseriidid, kreatiniin ja uuringus tuvastatud neli suremusega seotud biomarkerit) ning suremusriski mõjutavad olulised tausttunnused (näiteks sugu, vanus, kas indiviidil on diabeet, südame-veresoonkonnahaigusi või vähk). Tausttunnused koguti indiviidi geenidoonoriks hakkamise hetkel. Vaadeldavate tunnuste

jaotusest annab ülevaate tabel 4, kus kvantitatiivsete tunnuste korral on ära toodud iga tunnuse keskmine, standardhälve, miinimum ja maksimum. Nelja kvalitatiivse tunnuse (kõik on binaarsed väärtustega 0 ja 1) korral on tabelis 4 antud väärtuse 1 sagedus ning osakaal. Geenidoonori sugu tähistaval tunnusel on väärtus 1, kui indiviid on naissoost, ja väärtus 0, kui tegu on mehega. Ülejäänud kolme binaarse tunnuse korral näitab väärtus 1 vastavalt diabeedi, südame-veresoonkonnahaiguste või vähi olemasolu.

Tabel 4. Geenidoonorite ($n = 9787$) klasterdamiseks valitud tunnuste põhinäitajad

Tunnus (ühik)	Keskmine/ Sagedus (%)	Standardhälve	Min	Max
Sugu: naine	6306 (64,4%)	-	-	-
Vanus (aastat)	45,2	17,5	17	103
Suitsetamise kestus (aastat)	8,2	12,9	0	65
Diabeet	732 (7,5%)	-	-	-
Südame-veresoonkonnahaigus	887 (9,1%)	-	-	-
Vähk	358 (3,7%)	-	-	-
HDL-kolesterool (mmol/l)	1,69	0,37	0,37	4,21
Triglütseriidid (mmol/l)	1,54	0,99	0,07	9,88
Kreatiniin (μmol/l)	62,2	19,2	9,0	605,9
Albumiin (μmol/l)	100,4	7,1	55,1	135,6
α-1 glükoproteiin (mmol/l)	1,55	0,27	0,98	4,38
Tsitraat (μmol/l)	99,1	33,7	1,0	256,8
VLDL-partikli diameeter (nm)	37,0	1,9	32,1	45,1

Analüüsitav andmestik sisaldab 6306 naise ja 3481 mehe andmeid. Geenidoonorite vanus küsitlushetkel jäi 17 ja 103 eluaasta vahele, keskmine vanus oli 45,2 aastat. Diabeet on 732 geenidoonoril, 887 indiviidil oli diagnoositud mõni südame-veresoonkonnahaigus ning 358 oli vähidiagnoos. Kuna rohkem kui pooled geenidoonoritest ei olnud kunagi suitsetanud, on üle poolte suitsetamise kestuse tunnuse väärtustest nullid. Seetõttu teisendame selle tunnuse kvalitatiivseks tunnuseks. Uuel suitsetamise kestuse tunnusel on väärtus 1, kui geenidoonor polnud kunagi suitsetanud või oli suitsetanud vähem kui aasta aega (vastavalt 5652 ja 17 indiviidi); väärtus 2, kui geenidoonor oli suitsetanud üks kuni viis aastat (708 indiviidi); väärtus 3, kui geenidoonor oli suitsetanud kuus kuni 20 aastat (1638 indiviidi); ning väärtus 4, kui geenidoonor oli suitsetanud kauem kui 20 aastat (1772 indiviidi).

2.2 Klasteranalüüsi ülesannete formuleerimine

Artiklis Fischer jt (2014) toodi välja, et nelja suremusega seotud uue biomarkerita hinnatud Coxi võrdeliste riskide mudeli kohaselt mõjutas 25–74-aastaste geenidoonorite suremusriski nende sugu, HDL-kolesterooli tase, triglütseriidide tase, kreatiniini tase, kas indiviid suitsetas või mitte, suitsetatud aastate arv ja kolm haiguste indikaatorit. Geenidoonorite vanust küsitlushetkel kasutati Coxi mudelis ajaskaalana. Kui mudelisse lisati avastatud neli uut biomarkerit, ei omanud statistiliselt olulist mõju suremusriskile enam teised kolm markerit (HDL-kolesterool, triglütseriidid, kreatiniin). Nendest tulemustest lähtuvalt vaatleme kahte klasteranalüüsi ülesannet. Esimeses ülesandes teostame klasteranalüüsi ilma nelja uue biomarkerita. Teises ülesandes asendame HDL-kolesterooli, triglütseriidide ning kreatiniini tunnused esimesest ülesandest välja jäänud biomarkeritega. Mõlemas ülesandes kasutame Coxi mudelites olnud kahe suitsetamise tunnuse asemel eelmises alapeatükis defineeritud kvalitatiivset suitsetamise kestuse tunnust.

Klasteranalüüsi rakendamise eesmärk on välja selgitada, kas kirjeldatud tunnuste korral eristuvad erineva suremusriskiga indiviidide grupid. Saadud klastreid aitavad kirjeldada biomarkerite ja tausttunnuste jaotused klastrites. Huvi pakub, kas kõrgema suremusega klastrites on mõne biomarkeri väärtused ülejäänud klastritega võrreldes märkimisväärselt suuremad või väiksemad. Tervise Arengu Instituudi Surma põhjuste registrist on teada, et 2017. aastaks oli vaatluse all olevast 9787 geenidoonorist 925 surnud (473 naist, 452 meest). Geenidoonoriks hakkamise hetkel oli nende seas 238 diabeetikut, 283 oli diagnoositud mõni südame-veresoonkonnahaigus ning 106 oli vähk. Need arvud on olulised klasteranalüüsi tulemuste interpreteerimiseks.

Mõlema ülesande korral rakendame mudelipõhist klasteranalüüsi klastrite arvudega $K = 1, \ldots, 10$. Antud juhul ei teki suure valimimahu tõttu segujaotuste parameetrite hindamisel probleeme. Seetõttu hindame ainult kõige üldisema mudeli, mis segatüüpi tunnuste jaoks paketis „Rmixmod“ saadaval on. Selle segumudeli korral eeldatakse, et klasterdatavad vaatlused on sõltumatud realisatsioonid segujaotusest (17), kus iga f_k, $k = 1, \ldots, K$, on diagonaalse kovariatsioonimaatriksiga $\mathbf{\Sigma}_k = \lambda_k B_k$ ($\lambda_k = |\mathbf{\Sigma}_k|^{1/p}$ ja $|B_k| = 1$) normaaljaotuse tihedusfunktsioon (Langrognet jt, 2018). Paketis „Rmixmod“ tähistatakse seda mudelit $[\pi_k \varepsilon_k^{jh} \lambda_k B_k]$, kus π_k näitab, et komponentide kaaludele ei ole kitsendusi seatud. Iga segumudeli parameetrite hindamisel rakendame EM-algoritmi 20 korda, kus maksimaalne iteratsioonide arv on 1000 ning logaritmilise tõepärafunktsiooni suhtelise muutuse koondumiskriteeriumis (6) on $\varepsilon = 0{,}001$. Optimaalse klastrite arvu valimiseks kasutame asümptootilist ICL kriteeriumi (13). Klasteranalüüsi teostamiseks kirjutatud tarkvara R kood on toodud lisas 4.

2.3 Klasteranalüüsi tulemused

Ülesanne 1

Esimeses püstitatud ülesandes kasutame geenidoonorite klasterdamiseks tunnuseid sugu, vanus, suitsetamise kestus, HDL-kolesterool, triglütseriidid, kreatiniin ja kolme haiguste indikaatorit. Asümptootilise ICL kriteeriumi kohaselt on parim kaheksakomponendiline segumudel. Selle mudeliga saadud kaheksa klastri suurused on

$$\hat{n}_1 = 1216, \quad \hat{n}_2 = 419, \quad \hat{n}_3 = 1505, \quad \hat{n}_4 = 1699,$$
$$\hat{n}_5 = 824, \quad \hat{n}_6 = 834, \quad \hat{n}_7 = 2136, \quad \hat{n}_8 = 1154.$$

Klasterdamiseks kasutatud kvantitatiivsete tunnuste keskmised nendes klastrites on ära toodud tabelis 6 lk 44 ja kvalitatiivsete tunnuste sagedused igas klastris on toodud tabelis 7 lk 44. Tabelis 7 on ära toodud ka surnud indiviidide arv ja osakaal (edaspidi suremusrisk) igas klastris. Klastrite kirjeldamisel lähtume mainitud kahes tabelis toodud tulemustest.

Tabelist 7 lk 44 selgub, et esimese klastri moodustavad peamiselt mehed, 1109 meest ja 107 naist. Selles klastris on üks diabeetik, kellelgi polnud küsitlushetkel ühtegi südame-veresoonkonnahaigust ning viiel indiviidil oli vähidiagnoos. Antud klastrisse paigutatud 1216 indiviidist on ainult 14 ehk 1,2% surnud. Järelikult võib esimest klastrit kirjeldada kui madala suremusriskiga tervete (vaatluse all olevate haiguste seisukohalt) meeste gruppi. Teise klastrisse kuuluval 275 mehel ja 144 naisel on keskmiselt kõige kõrgem triglütseriidide ja kreatiniini tase ning keskmiselt kõige madalam HDL-kolesterooli tase (vt joonis 8 lk 41 ja tabel 6 lk 44). Selles klastris olevatest indiviididest 140 (33%) on diabeet, 114 (27%) oli diagnoositud mõni südame-veresoonkonnahaigus ning 38 (9%) oli vähidiagnoos. Tegu on huvipakkuva riskigrupiga: 129 ehk 30,8% klastri indiviididest on surnud ja suurem osa ekstreemsete biomarkerite väärtustega geenidoonoritest on selles klastris.

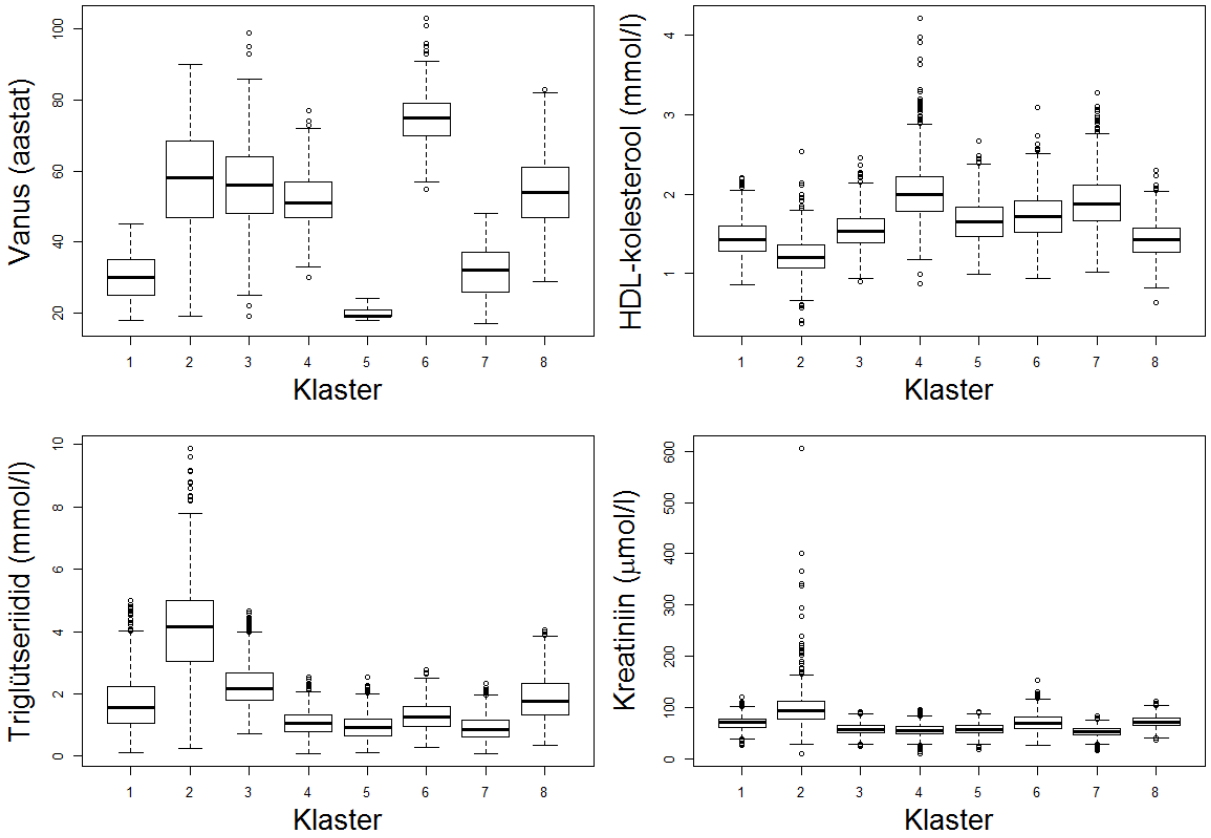
Kolmandasse klastrisse kuuluvad 1505 indiviidi on kõik naissoost. Neist 234 on diabeet, 148 oli geenidoonoriks hakkamise hetkel diagnoositud mõni südame-veresoonkonnahaigus ja 83 oli vähk. Suurem osa antud klastrisse paigutatud naistest ei ole kunagi suitsetanud. Seega moodustavad kolmanda klastri mittesuitsetavad ja haiged naised. Surnuid on selles klastris 151 (10%). Neljanda klastri moodustaval 1699 geenidoonoril (1415 naisel ja 284 mehel) on keskmiselt kõige kõrgem HDL-kolesterooli tase ja madal triglütseriidide ning kreatiniini tase. Võrreldes teiste klastritega, kus on samuti ülekaalukalt rohkem naisi, on sellele klastrile omapärane kaua suitsetanute suurem osakaal, ligi kolmandik neljanda klastri indiviididest on suitsetanud kauem kui 20 aastat.

Jooniselt 8 lk 41 on näha, et viienda klastri moodustavad noorimad geenidoonorid. Klastri

keskmine vanus on 19,7 aastat ja sarnaselt neljanda klastriga on antud klastris keskmiselt madal triglütseriidide tase. Kuuendasse klastrisse on paigutatud suurem osa vanematest geenidoonoritest, klastri keskmine vanus on 74,5 aastat. Tegu on loomuliku riskigrupiga, kus peamine riskitegur on vanus. Selles klastris on 207 geenidoonoril diabeet, 365 oli mõni südame-veresoonkonnahaigus ning 109 oli vähk, siia kuulub vastavalt 28%, 41% ja 30% kõigist haigetest. Enamik indiviididest selles klastris ei ole kunagi suitsetanud ning kolme vaatluse all oleva biomarkeri väärtused on normaaltasemel (biomarkerid ei paista silma eriliselt madalate või kõrgete väärtuste poolest). Ootuspäraselt on nendes kahes klastris vastavalt kõige madalam (0,2%) ja kõige kõrgem (39,5%) suremusrisk.

Seitsmenda klastri moodustavad peamiselt naissoost geenidoonorid (2099 naist, 37 meest), kellel on keskmiselt kõige madalam triglütseriidide ja kreatiniini tase ning kõrgemapoolne HDL-kolesterooli tase. Selles klastris on üks diabeetik, kolmel indiviidil oli diagnoositud mõni südame-veresoonkonnahaigus ning 17 indiviidil oli vähk. Vanusejaotuse ja surnute osakaalu poolest on antud klaster sarnane esimese ehk tervete meeste klastriga. Nendega võrreldes on naistel kõrgem HDL-kolesterooli tase, madalam triglütseriidide ja kreatiniini tase. Kokkuvõttes vastab seitsmes klaster madala suremusriskiga tervete naiste grupile. Viimasesse klastrisse paigutatud 1154 indiviidi on kõik meessoost. Neist rohkem kui pooled on suitsetanud kauem kui 20 aastat. Antud klastris on 126 diabeetikut, 227 indiviidil oli diagnoositud mõni südame-veresoonkonnahaigus ning 35 oli vähk. Kaheksandat klastrit võib seega kirjeldada kui kaua suitsetanud ja haigete meeste riskigruppi, 15,1% sellesse klastrisse kuuluvatest indiviididest on surnud. Biomarkerite väärtused selles klastris on sarnased esimese klastriga, kus on vähe haigeid, aga samuti palju kaua suitsetanud mehi (madalama keskmise vanuse tõttu jäi nende suitsetamise kestus vahemikku kuus kuni 20 aastat). Sarnast jagunemist võib täheldada neljanda ja seitsmenda klastri naiste korral.

Saadud kaheksa klastrit on hästi interpreteeritavad ning vastavad erineva suremusriskiga indiviidide gruppidele. Selgesti eristuvad mitmed meeste ja naiste grupid. Ootuspäraselt on surnute osakaal kõige väiksem noorimate geenidoonorite grupis ehk viiendas klastris ning kõige suurem kuuendas klastris, kuhu paigutati enamik vanematest geenidoonoritest. Huvipakkuva riskigrupi (teise klastri) moodustavad ekstreemsete biomarkerite väärtustega indiviidid. Selles klastris on keskmiselt kõige madalam HDL-kolesterooli tase ning keskmiselt kõige kõrgem triglütseriidide ja kreatiniini tase. Madal HDL-kolesterooli tase ja kõrge triglütseriidide tase suurendavad südame-veresoonkonnahaiguste tekkimise riski ning kõrgenenud kreatiniini tase võib viidata kroonilisele neeruhaigusele (Tartu Ülikooli Kliinikum, i.a). Järelikult on huvipakkuva riskigrupi ekstreemsed biomarkerite väärtused loogilised.



Joonis 8. Esimese ülesande klastrite kvantitatiivsete tunnuste karpdiagrammid

Ülesanne 2

Teises ülesandes on vaatluse all tunnused sugu, vanus, suitsetamise kestus, albumiin, α-1 glükoproteiin, tsitraat, VLDL-partikli diameeter ning kolm haiguste indikaatorit. Sel juhul on asümptootilise ICL kriteeriumi kohaselt parim seitsmekomponendiline segumudel. Selle mudeliga saadud klastrite mahud on järgmised:

$$\hat{n}_1 = 1468, \quad \hat{n}_2 = 961, \quad \hat{n}_3 = 1087, \quad \hat{n}_4 = 2082,$$
$$\hat{n}_5 = 1493, \quad \hat{n}_6 = 2113, \quad \hat{n}_7 = 583.$$

Järgnevalt kirjeldame neid klastreid tabelis 8 ja tabelis 9 lk 45 toodud tunnuste keskmiste ja sageduste abil.

Jooniselt 9 lk 43 on näha, et kõige vanemad geenidoonorid on paigutatud esimesse klastrisse. Antud klastri moodustava 1468 indiviidi hulgas 416 on diabeet, 589 oli geenidoonoriks hakkamise hetkel diagnoositud mõni südame-veresoonkonnahaigus ning 175 oli vähk (vastavalt 57%, 66% ja 49% kõigist haigetest). Selles klastris on keskmiselt kõige madalam albumiini tase ja kõrge tsitraadi tase. Teise klastri moodustavad 961 indiviidi on noorimad geenidoonorid, kellel on keskmiselt kõige kõrgem albumiini ning tsitraadi tase. Vanade geenidoonoritega võrreldes on noortel sarnane VLDL-partikli jaotus (vt

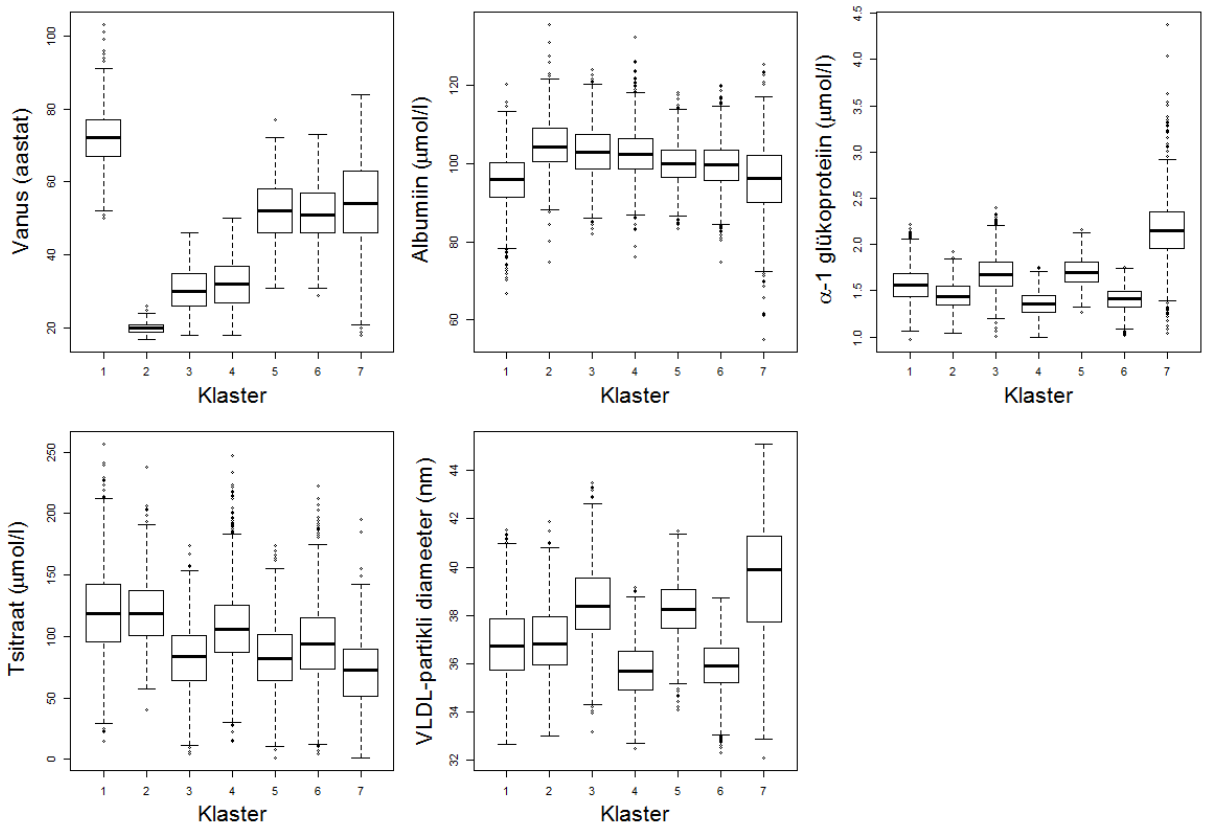
joonis 9 lk 43). Sarnaselt esimese ülesandega on ka siin nendes kahes klastris vastavalt kõige suurem (36,9%) ja kõige väiksem (0,3%) surnute osakaal.

Kolmandasse klastrisse paigutatud 1087 indiviidi (650 mehe, 437 naise) hulgas on kolm diabeetikut, kellelgi polnud küsitlushetkel ühtegi südame-veresoonkonnahaigust ning kolmel indiviidil oli vähk. Sellesse klastrisse kuuluvatest geenidoonoritest ligi pooled on suitsetanud kuus kuni 20 aastat. Neljandas klastris on ülekaalukalt rohkem naissoost geenidoonoreid (1692 naist ja 390 meest), kes vanusejaotuse poolest on sarnased eelmise klastriga. Antud klastris on samuti vähe haigeid ja märkimisväärne hulk kuus kuni 20 aastat suitsetanud indiviide. Kolmanda klastriga võrreldes on selles klastris madalam glükoproteiini tase, kõrgem tsitraadi tase ja väiksem VLDL-partikli diameeter. Mõlemas klastris on madal suremusrisk, kuna tegu on keskmiselt 30-aastaste tervete indiviididega. Nendes kahes klastris olevatest geenidoonoritest on vastavalt 1,2% ja 0,6% surnud.

Viienda ja kuuenda klastri moodustavad sarnase vanusejaotusega 1493 ja 2113 indiviidi, kellest ligikaudu kolmandik on suitsetanud üle 20 aasta. Kõrgema keskmise vanuse tõttu on eelmise kahe klastriga võrreldes nendes klastrites rohkem haigeid ja surnuid. Surnute osakaal on natuke suurem viiendas klastris, kus on rohkem haigeid, kõrgem α-1 glükoproteiini tase, madalam tsitraadi tase ja suurem VLDL-partikli diameeter. Seitsmenda klastri moodustava 583 indiviidi (401 mehe ja 182 naise) hulgast 115 (20%) on diabeet, 146 (25%) oli diagnoositud mõni südame-veresoonkonnahaigus ja 41 (7%) oli vähk. Ligi pooled antud klastri geenidoonoritest on suitsetanud kauem kui 20 aastat. Selles klastris on keskmiselt madal albumiini tase, keskmiselt kõige kõrgem α-1 glükoproteiini tase, kõige madalam tsitraadi tase ja kõige suurem VLDL-partikli diameeter. Seega vastab viimane klaster huvipakkuvale riskigrupile, surnuid on selles klastris 143 (24,5%).

Teise ülesande kahe suurima surnute osakaaluga klastris on keskmiselt kõige madalam albumiini tase ning kolme kõige madalama suremusriskiga klastris on keskmiselt kõige kõrgem albumiini tase. See on kooskõlas artiklis Fischer jt (2014) esitatud tulemustega, et albumiini madal tase suurendab suremusriski. Seitsmendas klastris ehk huvipakkuvas riskigrupis on lisaks keskmiselt kõige madalam tsitraadi tase, kõrgeim glükoproteiini tase ja suurim VLDL-partikli diameeter. Huvipakkuva riskigrupi keskmiselt kõige kõrgem α-1 glükoproteiini tase on samuti nimetatud artikli tulemustega kooskõlas (suurem väärtus suurendab riski). Üllatav on selle klastri keskmiselt kõige madalam tsitraadi tase ning keskmiselt kõige suurem VLDL-partikli diameeter, sest artikli Fischer jt (2014) kohaselt peaksid sellised väärtused suremusriski vähendama.

Vaadeldud kahes ülesandes saadud klasterduste risttabel on toodud tabelis 5 lk 43. Sellest tabelist on näha, et mõlema ülesande kõige nooremate geenidoonorite (vastavalt viies ja teine) klaster ning tervete naiste (vastavalt seitsmes ja neljas) klaster koosnevad valdavalt



Joonis 9. Teise ülesande klastrite kvantitatiivsete tunnuste karpdiagrammid

samadest indiviididest. Teise ülesande esimene ehk vanade geenidoonorite klaster sisaldab rohkem kui 90% esimese ülesande vastava klastri indiviididest, aga lisaks ka ligi neljandiku esimese ülesande huvipakkuvast riskigrupist (teine klaster) ja palju teistes klastrites olnud vanadest ning haigetest geenidoonoritest (vt ka tabel 7 lk 44 ja tabel 9 lk 45). Ülejäänud klastrite vahel nii suurt kokkulangevust ei ole. Saadud kahe huvipakkuva riskigrupi (teise ja seitsmenda klastri) ühisosa moodustavad 253 geenidoonorit, kellest 65 on surnud. Seda ühisosa võib vaadelda omaette riskigrupina.

Tabel 5. Esimese ülesande kaheksa ja teise ülesande seitsme klastri risttabel

Klaster (maht)	1 (1468)	2 (961)	3 (1087)	4 (2082)	5 (1493)	6 (2113)	7 (583)
1 (1216)	0	103	687	363	13	24	26
2 (419)	97	0	14	2	51	2	253
3 (1505)	318	1	74	15	736	281	80
4 (1699)	86	0	9	46	203	1329	26
5 (824)	0	760	33	29	0	0	2
6 (834)	771	0	0	0	12	36	15
7 (2136)	0	97	266	1626	32	113	2
8 (1154)	196	0	4	1	446	328	179

Tabel 6. Esimese ülesande kvantitatiivsete tunnuste keskmised saadud klastrites

Tunnus (ühik) \ Klaster (maht)	1 (1216)	2 (419)	3 (1505)	4 (1699)	5 (824)	6 (834)	7 (2136)	8 (1154)
Vanus (aastat)	29,9	57,3	56,0	52,1	19,7	74,5	31,8	54,7
HDL-kolesterool (mmol/l)	1,45	1,22	1,54	2,03	1,66	1,72	1,90	1,43
Triglütseriidid (mmol/l)	1,72	4,14	2,28	1,07	0,96	1,29	0,92	1,87
Kreatiniin (μmol/l)	69,9	100,6	57,8	55,2	57,4	71,1	52,0	71,8

Tabel 7. Kvalitatiivsete tunnuste sagedused esimeses ülesandes saadud klastrites

Tunnus (väärtus) \ Klaster (maht)	1 (1216)	2 (419)	3 (1505)	4 (1699)	5 (824)	6 (834)	7 (2136)	8 (1154)
Mees	1109	275	0	284	337	285	37	1154
Naine	107	144	1505	1415	487	549	2099	0
Suitsetamise kestus ($<$ 1 a)	536	188	1129	955	498	656	1337	370
Suitsetamise kestus (1–5 a)	135	18	17	28	324	15	153	18
Suitsetamise kestus (6–20 a)	545	68	116	104	2	38	646	119
Suitsetamise kestus ($>$ 20 a)	0	145	243	612	0	125	0	647
Diabeet	1	140	234	21	2	207	1	126
Südame-veresoonkonnahaigus	0	114	148	30	0	365	3	227
Vähk	5	38	83	60	11	109	17	35
Surnud	14 (1,2%)	129 (30,8%)	151 (10,0%)	107 (6,3%)	2 (0,2%)	329 (39,5%)	19 (0,9%)	174 (15,1%)

Tabel 8. Teise ülesande kvantitatiivsete tunnuste keskmised saadud klastrites

Tunnus (ühik) \ Klaster (maht)	1 (1468)	2 (961)	3 (1087)	4 (2082)	5 (1493)	6 (2113)	7 (583)
Vanus (aastat)	72,1	20,0	30,1	31,9	52,3	51,6	53,8
Albumiin (μmol/l)	95,6	104,8	103,1	102,6	100,1	99,7	95,8
α-1 glükoproteiin (mmol/l)	1,57	1,44	1,69	1,36	1,71	1,41	2,18
Tsitraat (μmol/l)	119,8	120,1	83,1	107,4	82,6	94,7	71,3
VLDL-partikli diameeter (nm)	36,8	37,0	38,5	35,7	38,3	35,9	39,4

Tabel 9. Kvalitatiivsete tunnuste sagedused teises ülesandes saadud klastrites

Tunnus (väärtus) \ Klaster (maht)	1 (1468)	2 (961)	3 (1087)	4 (2082)	5 (1493)	6 (2113)	7 (583)
Mees	527	413	650	390	518	582	401
Naine	941	548	437	1692	975	1531	182
Suitsetamise kestus ($<$ 1 a)	1104	591	460	1271	805	1233	205
Suitsetamise kestus (1–5 a)	24	362	119	126	21	41	15
Suitsetamise kestus (6–20 a)	71	8	508	685	138	148	80
Suitsetamise kestus ($>$ 20 a)	269	0	0	0	529	691	283
Diabeet	416	2	3	0	159	37	115
Südame-veresoonkonnahaigus	589	0	0	1	80	71	146
Vähk	175	11	3	15	47	66	41
Surnud	541 (36,9%)	3 (0,3%)	13 (1,2%)	13 (0,6%)	104 (7,0%)	108 (5,1%)	143 (24,5%)

Kokkuvõte

Antud magistritöö eesmärk oli anda ülevaade mudelipõhise klasteranalüüsi teostamisest kvantitatiivsete, kvalitatiivsete ja segatüüpi tunnuste korral. Töö esimeses osas selgitati, kuidas mitmemõõtmeliste normaaljaotuste segu ja multinomiaalsete jaotuste segu abil vastavalt kvantitatiivseid ja kvalitatiivseid andmeid klasterdatakse ning kuidas mudelipõhise lähenemise korral segatüüpi tunnuseid käsitletakse. Põhjalikult kirjeldati, kuidas EM-algoritmiga eeldatavate segujaotuste parameetrite hinnangud leitakse ning kuidas parameetrite arvu kitsenduste seadmisega vähendatakse. Lisaks tuletati integreeritud klassifitseerimistõepära ehk ICL kriteerium, mida kasutatakse erinevate kitsenduste ja komponentide arvuga hinnatud segumudelite hulgast parima mudeli valimiseks.

Mudelipõhise klasteranalüüsi käitumise illustreerimiseks esitati kaks simulatsiooninäidet. Esimeses näites klasterdati erineva klastrite kattuvusega kvalitatiivseid andmeid. Kuna kvalitatiivsete tunnuste korral saab ICL kriteeriumi väärtuse täpselt välja arvutada, võrreldi selles näites ka asümptootilise ja täpse ICL kriteeriumi kohaselt parimate segumudelite tulemusi. Saadud tulemused kinnitasid, et kvalitatiivsete tunnuste korral on segumudeli kuju ja klastrite arvu valimiseks parem kasutada täpset ICL kriteeriumi. Teises simulatsiooninäites võrreldi populaarse kaugusi kasutava K-keskmiste meetodi ja mudelipõhise klasteranalüüsi tulemusi erineva kujuga klastrite korral. Selles näites leidis kinnitust, et normaaljaotuste segust genereeritud vaatluste korral töötab K-keskmiste meetod mudelipõhise lähenemisega võrreldes sama hästi ainult siis, kui klastrid on sfäärilised ja üksteisest selgesti eraldatud.

Töö teises osas rakendati mudelipõhist klasterdamist Tartu Ülikooli Eesti Geenivaramu segatüüpi andmetele. Vaatluse all oli kaks klasteranalüüsi ülesannet, kus 9787 geenidoonori grupeerimiseks kasutati erinevaid biomarkereid ja olulisi suremust mõjutavaid tausttunnuseid. Eesmärk oli välja selgitada, kas kahes püstitatud ülesandes eristuvad erineva suremusriskiga indiviidide grupid. Klastrite interpreteerimiseks kasutati Tervise Arengu Instituudi Surma põhjuste registrist saadud surnud geenidoonorite arvu. Saadud klastrid vastasid erineva suremusriskiga indiviidide gruppidele. Esimeses ülesandes, kus artiklis Fischer jt (2014) tuvastatud neli suremusega seotud uut markerit jäeti kõrvale, eristusid mitmed meeste ja naiste grupid. Kui klasterdamiseks kasutati kõrvale jäänud biomarkereid ehk albumiini, α-1 glükoproteiini, tsitraati ja VLDL-partikli diameetrit, oli ainult kahes klastris ülekaalukalt rohkem ühest soost indiviide. Mõlemas ülesandes tekkis noorte ja vanade geenidoonorite grupp, kus ootuspäraselt oli vastavalt kõige madalam ja kõige kõrgem suremusrisk. Samuti eristus mõlemas ülesandes ekstreemsete biomarkerite väärtustega klaster, kus oli ka suur surnute osakaal.

Kasutatud kirjandus

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. doi: 10.1109/tac.1974.1100705

Allik, A. (2014). *Eesti ja Soome teadlaste ühisuuring avastas neli suremusega seotud biomarkerit.* Tartu Ülikooli genoomika instituudi koduleheküIg. Kasutatud 07.05.2019, `http://www.geenivaramu.ee/et/uudised/eesti-soome-teadlaste-uhisuuring-avastas-neli-suremusega-seotud-biomarkerit`

Banfield, J. D. ja Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometric, 49*(3), 803–821. doi: 10.2307/2532201

Biernacki, C., Celeux, G. ja Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(7), 719–725. doi: 10.1109/34.865189

Biernacki, C., Celeux, G. ja Govaert, G. (2010). Exact and Monte Carlo calculations of integrated likelihoods for the latent class model. *Journal of Statistical Planning and Inference, 140*(11), 2991–3002. doi: 10.1016/j.jspi.2010.03.042

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer, New York.

Celeux, G. ja Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis, 14*(3), 315–332. doi: 10.1016/0167-9473(92)90042-E

Fischer, K., Kettunen, J., Würtz, P., Haller, T., Havulinna, A. S., Kangas, A. J. jt (2014). Biomarker profiling by nuclear magnetic resonance spectroscopy for the prediction of all-cause mortality: an observation study of 17,345 persons. *PLOS Medicine, 11*(2), e1001606. doi: 10.1371/journal.pmed.1001606

Foss, A. H. ja Markatou, M. (2018). kamila: clustering mixed-type data in R and Hadoop. *Journal of Statistical Software, 83*(13), 1–44. doi: 10.18637/jss.v083.i13

Hastie, T., Tibshirani, R. ja Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York.

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

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

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

Langrognet, F., Lebret, R., Poli, C., Iovleff, S., Auder, B., Bhatia, P. jt (2018). Package 'Rmixmod'. *Classification with mixture modelling*. Kasutatud 03.05.2019, `https://cran.r-project.org/web/packages/Rmixmod/Rmixmod.pdf`

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. doi: 10.1214/aos/1176344136

Tartu Ülikooli Kliinikum. (i.a). *Ühendlabori käsiraamat*. Kasutatud 07.05.2019, `https://www.kliinikum.ee/yhendlabor/analueueside-taehestikuline-register`

Tibshirani, R., Walther, G. ja Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411–423. doi: 10.1111/1467-9868.00293

Lisad

Lisa 1. Jooniste 1 ja 2 R-kood

```
# Funktsioon, mis joonistab hajuvusellipsi kanoonilise võrrandi põhjal
ellips <- function(mu,Sigma,const) {
  eig=eigen(Sigma)
  lambda=diag(eig$values)
  gamma=eig$vectors
  theta=seq(from=0,to=2*pi,length=100)
  koords=matrix(NA,nrow=2,ncol=100)
  for (i in 1:100) {
    koords[,i]=mu+sqrt(const)*gamma%*%lambda^(1/2)%*%rbind(cos(theta[i]),sin(theta[i]))
  }
  lines(x=koords[1,],y=koords[2,],col="black",type="l")
}


# Joonis 1 ----
sigma1=matrix(c(3,4,4,9),nrow=2,ncol=2); det(sigma1)
eig1=eigen(sigma1)
lambda1=det(sigma1)^(1/2)
D1=eig1$vectors
A1=diag(eig1$values/lambda1)
lambda1;D1;A1

sigma2=matrix(c(9.79,1.98,1.98,12.76)/sqrt(11),nrow=2,ncol=2); det(sigma2)
eig2=eigen(sigma2)
lambda2=det(sigma2)^(1/2)
D2=eig2$vectors
A2=diag(eig2$values/lambda2)
lambda2;D2;A2

sigma3=2*sigma2; det(sigma3)
eig3=eigen(sigma3)
lambda3=det(sigma3)^(1/2)
D3=eig3$vectors
A3=diag(eig3$values/lambda3)
lambda3;D3;A3

vals=c(0.1,0.2,0.3,0.4)
plot(0,type="n",xlim=c(-4,8),ylim=c(-2.5,2.5),xaxt="n",yaxt="n",xlab="",ylab="")
for (i in 1:length(vals)) {
  ellips(mu=c(-2,0),Sigma=sigma1,const=vals[i])
  ellips(mu=c(2,0),Sigma=sigma2,const=vals[i])
  ellips(mu=c(6,0),Sigma=2*sigma2,const=vals[i])
}


# Joonis 2 ----
par(mfrow=c(2,2),mar=c(0,0.5,5,0.5),oma=c(1,0,0,0),cex.main=3.5)
limit=c(-4,4)

# 1. Sigma_k = lambda * I_2
mu1=c(-2,0); mu2=c(2,0); sigma=diag(2)
```

```
plot(0,type="n",xlim=limit,ylim=limit,xaxt="n",yaxt="n",xlab="",ylab="")
title(main=expression(paste(Sigma[k],"=",lambda,"I"[2])),line=2.5)
ellips(mu=mu1,Sigma=sigma,const=1)
ellips(mu=mu2,Sigma=sigma,const=1)

# 2. Sigma_k = lambda_k * I_2
plot(0,type="n",xlim=limit,ylim=limit,xaxt="n",yaxt="n",xlab="",ylab="")
title(main=expression(paste(Sigma[k],"=",lambda[k],"I"[2])),line=2.5)
ellips(mu=mu1,Sigma=sigma,const=1)
ellips(mu=mu2,Sigma=3*sigma,const=1)

# 3. Sigma_k = lambda_k * D * A * D'
sigma2=matrix(c(3,4,4,9),nrow=2,ncol=2)
plot(0,type="n",xlim=limit,ylim=limit,xaxt="n",yaxt="n",xlab="",ylab="")
title(main=expression(paste(Sigma[k],"=",lambda[k],"DAD'")),line=2.5)
ellips(mu=mu1,Sigma=sigma2,const=0.4)
ellips(mu=mu2,Sigma=3*sigma2,const=0.4)

# 4. Sigma_k = lambda_k * D_k * A * (D_k)'
sigma3=matrix(c(3,-4,-4,9),nrow=2,ncol=2)
plot(0,type="n",xlim=limit,ylim=limit,xaxt="n",yaxt="n",xlab="",ylab="")
title(main=expression(paste(Sigma[k],"=",lambda[k],D[k],"A","D'"[k])),line=2.5)
ellips(mu=mu1,Sigma=sigma2,const=0.4)
ellips(mu=mu2,Sigma=3*sigma3,const=0.4)
```

Lisa 2. Näite 2 R-kood

```
library(Rmixmod)

# Funktsioon, mis teostab kvalitatiivsete andmete mudelipõhise klasteranalüüsi
klasterda1 <- function(andmed) {
  # Vaatluse all on 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=20,nbIterationInAlgo=1000),
                       seed=12,criterion="ICL")
}

# Funktsioon, mis teeb Rmixmod tulemuste andmestiku
mixmod_tulem1 <- function(valim,mudelid) {
  tulemused=data.frame()
  for (i in 1:length(mudelid)) {
    tulemused[i,1]=mudelid[[i]]@model
    tulemused[i,2]=mudelid[[i]]@nbCluster
    tulemused[i,3]=mudelid[[i]]@likelihood
    tulemused[i,4]=mudelid[[i]]@criterionValue # ICL
    tulemused[i,5]=ICLtapne(valim,mudelid[[i]],m=c(3,3,4,4)) # ICL_M
  }
  colnames(tulemused)=c("Mudel","K","Toepara","ICL","ICLtapne")
  return(tulemused)
}

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

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

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

# 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)
```

```
}

# Funktsioon, mis genereerib valimi kahe komponendiga multinomiaalsete jaotuste segust
genAndmed <- function(n,pi,delta) {
  # Multinomiaalsete jaotuste tõenäosuste arvutamine
  alfa1=alfa2=rep(NA,times=sum(m)) # K=2
  loendur=0
  for (j in 1:p) { # p=4
    for (h in 1:m[j]) { # m1=m2=3 ja m3=m4=4
      loendur=loendur+1
      alfa1[loendur]=alfa(k=1,j=j,h=h,delta=delta)
      alfa2[loendur]=alfa(k=2,j=j,h=h,delta=delta)
    }
  }

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

  # Esimene klaster
  x1a=data.frame(t(rmultinom(n=n1,size=1,prob=alfa1[1:3])))
  x1b=data.frame(t(rmultinom(n=n1,size=1,prob=alfa1[4:6])))
  x1c=data.frame(t(rmultinom(n=n1,size=1,prob=alfa1[7:10])))
  x1d=data.frame(t(rmultinom(n=n1,size=1,prob=alfa1[11:14])))
  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[1:3])))
  x2b=data.frame(t(rmultinom(n=n2,size=1,prob=alfa2[4:6])))
  x2c=data.frame(t(rmultinom(n=n2,size=1,prob=alfa2[7:10])))
  x2d=data.frame(t(rmultinom(n=n2,size=1,prob=alfa2[11:14])))
  x2a2=data.frame("K2"=apply(x2a,1,function(x) which(x==max(x))))
  x2b2=data.frame("K2"=apply(x2b,1,function(x) which(x==max(x))))
  x2c2=data.frame("K2"=apply(x2c,1,function(x) which(x==max(x))))
  x2d2=data.frame("K2"=apply(x2d,1,function(x) which(x==max(x))))
  x2=cbind(x2a2,x2b2,x2c2,x2d2,2)
  colnames(x2)=c("X1","X2","X3","X4","K")

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


# Klasterdatavate vaatluste genereerimine ----
K=2; pi=c(0.3,0.7)
p=4; m=c(3,3,4,4)

# Situatsioon 1
valimA1=genAndmed(n=1000,pi,delta=0.4)
```

```
valimA2=genAndmed(n=3000,pi,delta=0.4)

# Situatsioon 2
valimB1=genAndmed(n=1000,pi,delta=0.6)
valimB2=genAndmed(n=3000,pi,delta=0.6)


# Mudelipõhine klasteranalüüs ----

# Situatsioon 1a (n=1000)
mudelA1=klasterda1(andmed=valimA1[,-5])
(tulemusedA1=mixmod_tulem1(valim=valimA1,mudelid=mudelA1@results))
(parimA11=mudelA1["results"][[which.min(tulemusedA1$ICL)]]) # [e]
table(valimA1$K,parimA11@partition)
(parimA12=mudelA1["results"][[3]]) # parim ICLtapne kohaselt on [e^j]
table(valimA1$K,parimA12@partition)

# Situatsioon 1b (n=3000)
mudelA2=klasterda1(andmed=valimA2[,-5])
(tulemusedA2=mixmod_tulem1(valim=valimA2,mudelid=mudelA2@results))
(parimA21=mudelA2["results"][[which.min(tulemusedA2$ICL)]]) # [e^j]
table(valimA2$K,parimA21@partition)
(parimA22=mudelA2["results"][[2]]) # parim ICLtapne kohaselt on [e]
table(valimA2$K,parimA22@partition)

# Situatsioon 2a (n=1000)
mudelB1=klasterda1(andmed=valimB1[,-5])
(tulemusedB1=mixmod_tulem1(valim=valimB1,mudelid=mudelB1@results))
(parimB11=mudelB1["results"][[4]]) # [e] kahe komponendiga
table(valimB1$K,parimB11@partition)
(parimB12=mudelB1["results"][[5]]) # [e^j] kahe komponendiga
table(valimB1$K,parimB12@partition)

# Situatsioon 2b (n=3000)
mudelB2=klasterda1(andmed=valimB2[,-5])
(tulemusedB2=mixmod_tulem1(valim=valimB2,mudelid=mudelB2@results))
(parimB21=mudelB2["results"][[4]]) # [e^j] kahe komponendiga
table(valimB2$K,parimB21@partition)
(parimB22=mudelB2["results"][[5]]) # [e] kahe komponendiga
table(valimB2$K,parimB22@partition)

# Joonis 3 ----
maxA1=apply(parimA12@proba,1,FUN=max)
maxA2=apply(parimA21@proba,1,FUN=max)
maxB1=apply(parimB12@proba,1,FUN=max)
maxB2=apply(parimB21@proba,1,FUN=max)

par(mfrow=c(2,2),mar=c(5,5,2,0.5),cex.main=1.5,cex.lab=1.5)
hist(maxA1,labels=TRUE,ylim=c(0,1000),breaks=seq(0.5,1,0.1),main="Situatsioon 1 (n=1000)",
     xlab=expression(paste("max ",hat(gamma)[i])),ylab="Sagedus")
hist(maxB1,labels=TRUE,ylim=c(0,1000),breaks=seq(0.5,1,0.1),main="Situatsioon 2 (n=1000)",
     xlab=expression(paste("max ",hat(gamma)[i])),ylab="Sagedus")
hist(maxA2,labels=TRUE,ylim=c(0,3000),breaks=seq(0.5,1,0.1),main="Situatsioon 1 (n=3000)",
     xlab=expression(paste("max ",hat(gamma)[i])),ylab="Sagedus")
hist(maxB2,labels=TRUE,ylim=c(0,3000),breaks=seq(0.5,1,0.1),main="Situatsioon 2 (n=3000)",
     xlab=expression(paste("max ",hat(gamma)[i])),ylab="Sagedus")
par(mfrow=c(1,1))
```

Lisa 3. Näite 3 R-kood

```
library(Rmixmod)
library(mvtnorm)
library(cluster)

# Funktsioon, mis teeb Rmixmod tulemuste andmestiku
mixmod_tulem2 <- function(mudelid) {
  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
  }
  colnames(tulemused)=c("Mudel","K","Toepara","ICL")
  return(tulemused)
}

# Funktsioon, mis teostab mudelipõhise klasteranalüüsi ja tagastab parima mudeli tulemused
klasterda2 <- function(andmed,k) {
  mixmod=mixmodCluster(data=data.frame(andmed),nbCluster=k,
                       dataType="quantitative",
                       models=mixmodGaussianModel(),
                       strategy=mixmodStrategy(nbTry=20,nbIterationInAlgo=1000),
                       seed=12,criterion="ICL")
  tulemused=mixmod_tulem2(mixmod["results"]) # kõik tulemused
  nrow=min(which(tulemused$Toepara != 0)) # esimene (parim) mudel, kus parameetrid hinnatud
  mixmod["results"][[nrow]] # vastav mudel
}

# Funktsioon, mis joonistab kõigi klasterduste joonised
joonised <- function(x) {
  par(mfrow=c(2,5),pty="s",mar=c(0,0,0,0),mgp=c(0,0.5,0),mai=c(0.1,0.2,0.2,0.1),cex.main=1.5)
  plot(x,xlab="",ylab="",col=kx2$cluster+1,main="K-keskmine, K=2",asp=1,xaxt="n")
  axis(1,labels=FALSE)
  points(kx2$center,col=1,pch=22,bg=2:3,lwd=2,cex=2)
  plot(x,xlab="",ylab="",col=kx3$cluster+1,main="K-keskmine, K=3",asp=1,xaxt="n",yaxt="n")
  axis(1,labels=FALSE); axis(2,labels=FALSE)
  points(kx3$center,col=1,pch=22,bg=2:4,lwd=2,cex=2)
  plot(x,xlab="",ylab="",col=kx4$cluster+1,main="K-keskmine, K=4",asp=1,xaxt="n",yaxt="n")
  axis(1,labels=FALSE); axis(2,labels=FALSE)
  points(kx4$center,col=1,pch=22,bg=2:5,lwd=2,cex=2)
  plot(x,xlab="",ylab="",col=kx5$cluster+1,main="K-keskmine, K=5",asp=1,xaxt="n",yaxt="n")
  points(kx5$center,col=1,pch=22,bg=2:6,lwd=2,cex=2)
  axis(1,labels=FALSE); axis(2,labels=FALSE)
  plot(x,xlab="",ylab="",col=kx6$cluster+1,main="K-keskmine, K=6",asp=1,xaxt="n",yaxt="n")
  axis(1,labels=FALSE); axis(2,labels=FALSE)
  points(kx6$center,col=1,pch=22,bg=2:7,lwd=2,cex=2)
  plot(x,xlab="",ylab="",col=mudel2@partition+1,main="Mudelipõhine, K=2",asp=1)
  axis(1,labels=FALSE)
  points(mudel2@parameters@mean,col=1,pch=22,bg=2:3,lwd=2,cex=2)
  plot(x,xlab="",ylab="",col=mudel3@partition+1,main="Mudelipõhine, K=3",asp=1,yaxt="n")
  axis(1,labels=FALSE); axis(2,labels=FALSE)
  points(mudel3@parameters@mean,col=1,pch=22,bg=2:4,lwd=2,cex=2)
  plot(x,xlab="",ylab="",col=mudel4@partition+1,main="Mudelipõhine, K=4",asp=1,yaxt="n")
  axis(1,labels=FALSE); axis(2,labels=FALSE)
```

```
  points(mudel4@parameters@mean,col=1,pch=22,bg=2:5,lwd=2,cex=2)
  plot(x,xlab="",ylab="",col=mudel5@partition+1,main="Mudelipõhine, K=5",asp=1,yaxt="n")
  axis(1,labels=FALSE); axis(2,labels=FALSE)
  points(mudel5@parameters@mean,col=1,pch=22,bg=2:6,lwd=2,cex=2)
  plot(x,xlab="",ylab="",col=mudel6@partition+1,main="Mudelipõhine, K=6",asp=1,yaxt="n")
  axis(2,labels=FALSE)
  points(mudel6@parameters@mean,col=1,pch=22,bg=2:7,lwd=2,cex=2)
  par(mfrow=c(1,1))
}

# Andmete genereerimine ----
# Situatsioon 1
set.seed(12)
mu1=c(5,0); mu2=c(35,20); mu3=c(35,-20)
sigma=20*diag(2)
komp1=rmvnorm(n=200,mu1,sigma)
komp2=rmvnorm(n=200,mu2,sigma)
komp3=rmvnorm(n=200,mu3,sigma)
x1=rbind(komp1,komp2,komp3)

# Situatsioon 2
set.seed(12)
mu1=c(5,5); mu2=c(40,10); mu3=c(35,-20)
sigma1=matrix(c(35,5,5,75),nrow=2,ncol=2,byrow=TRUE)
sigma2=matrix(c(65,-25,-25,30),nrow=2,ncol=2,byrow=TRUE)
sigma3=matrix(c(250,15,15,40),nrow=2,ncol=2,byrow=TRUE)
komp1=rmvnorm(n=100,mu1,sigma1)
komp2=rmvnorm(n=200,mu2,sigma2)
komp3=rmvnorm(n=300,mu3,sigma3)
x2=rbind(komp1,komp2,komp3)

# Situatsioon 3
set.seed(12)
mu1=c(5,0); mu2=c(40,45); mu3=c(35,-20)
sigma1=matrix(c(1,-10,-10,400),nrow=2,ncol=2,byrow=TRUE)
sigma2=matrix(c(45,5,5,1),nrow=2,ncol=2,byrow=TRUE)
sigma3=matrix(c(5,1,1,235),nrow=2,ncol=2,byrow=TRUE)
komp1=rmvnorm(n=200,mu1,sigma1)
komp2=rmvnorm(n=200,mu2,sigma2)
komp3=rmvnorm(n=200,mu3,sigma3)
x3=rbind(komp1,komp2,komp3)

# Joonis 4 ----
par(mfrow=c(1,3),pty="s",mar=c(0,0,0,0),mgp=c(0,0.5,0),mai=c(0.1,0.2,0.4,0.1))
varv1=c(rep(1,200),rep(2,200),rep(3,200))
varv2=c(rep(1,100),rep(2,200),rep(3,300))
plot(x1,xlab="",ylab="",col=varv1,main="Situatsioon 1",asp=1,cex.main=2)
plot(x2,xlab="",ylab="",col=varv2,main="Situatsioon 2",asp=1,cex.main=2)
plot(x3,xlab="",ylab="",col=varv1,main="Situatsioon 3",asp=1,cex.main=2)
par(mfrow=c(1,1))


# Mudelipõhine klasteranalüüs ----

# Situatsioon 1 ----
mudel2=klasterda2(andmed=x1,k=2)
mudel3=klasterda2(andmed=x1,k=3)
```

```
mudel4=klasterda2(andmed=x1,k=4)
mudel5=klasterda2(andmed=x1,k=5)
mudel6=klasterda2(andmed=x1,k=6)

kx2=kmeans(x1,centers=2,iter.max=100,nstart=50)
kx3=kmeans(x1,centers=3,iter.max=100,nstart=50)
kx4=kmeans(x1,centers=4,iter.max=100,nstart=50)
kx5=kmeans(x1,centers=5,iter.max=100,nstart=50)
kx6=kmeans(x1,centers=6,iter.max=100,nstart=50)

gapK=clusGap(x1,kmeans,nstart=50,K.max=6,B=600,spaceH0="original")
print(gapK,method="Tibs2001SEmax") # K=3

# K-keskmiste meetodi tulemuste andmestik
tulemused1=data.frame("Meetod"=rep("K-keskmine",5),"K"=c(2:6))
kx=list(kx2,kx3,kx4,kx5,kx6)
tulemused1$ASW=lapply(X=kx,FUN=function(res)
  summary(silhouette(res$cluster,dist(x1)^2))$avg.width)

# Mudelipõhise klasteranalüüsi tulemuste andmestik
tulemused2=data.frame("Meetod"=rep("Mudelipõhine",5),"K"=c(2:6))
mudelid=list(mudel2,mudel3,mudel4,mudel5,mudel6)
tulemused2$Mudel=lapply(X=mudelid,FUN=function(res) res@model)
tulemused2$Toepara=lapply(X=mudelid,FUN=function(res) res@likelihood)
tulemused2$ICL=lapply(X=mudelid,FUN=function(res) res@criterionValue[1])
tulemused2$ASW=lapply(X=mudelid,FUN=function(res)
  summary(silhouette(res@partition,dist(x1)^2))$avg.width)

# Joonis 5 ----
joonised(x=x1)


# Situatsioon 2 ----
mudel2=klasterda2(andmed=x2,k=2)
mudel3=klasterda2(andmed=x2,k=3)
mudel4=klasterda2(andmed=x2,k=4)
mudel5=klasterda2(andmed=x2,k=5)
mudel6=klasterda2(andmed=x2,k=6)

kx2=kmeans(x2,centers=2,iter.max=100,nstart=50)
kx3=kmeans(x2,centers=3,iter.max=100,nstart=50)
kx4=kmeans(x2,centers=4,iter.max=100,nstart=50)
kx5=kmeans(x2,centers=5,iter.max=100,nstart=50)
kx6=kmeans(x2,centers=6,iter.max=100,nstart=50)

gapK=clusGap(x2,kmeans,nstart=50,K.max=6,B=600,spaceH0="original")
print(gapK,method="Tibs2001SEmax") # K=1

# K-keskmiste meetodi tulemuste andmestik
tulemused1=data.frame("Meetod"=rep("K-keskmine",5),"K"=c(2:6))
kx=list(kx2,kx3,kx4,kx5,kx6)
tulemused1$ASW=lapply(X=kx,FUN=function(res)
  summary(silhouette(res$cluster,dist(x2)^2))$avg.width)

# Mudelipõhise klasteranalüüsi tulemuste andmestik
tulemused2=data.frame("Meetod"=rep("Mudelipõhine",5),"K"=c(2:6))
mudelid=list(mudel2,mudel3,mudel4,mudel5,mudel6)
```

```
tulemused2$Mudel=lapply(X=mudelid,FUN=function(res) res@model)
tulemused2$Toepara=lapply(X=mudelid,FUN=function(res) res@likelihood)
tulemused2$ICL=lapply(X=mudelid,FUN=function(res) res@criterionValue)
tulemused2$ASW=lapply(X=mudelid,FUN=function(res)
  summary(silhouette(res@partition,dist(x2)^2))$avg.width)

# Joonis 6 ----
joonised(x=x2)


# Situatsioon 3 ----
mudel2=klasterda2(andmed=x3,k=2)
mudel3=klasterda2(andmed=x3,k=3)
mudel4=klasterda2(andmed=x3,k=4)
mudel5=klasterda2(andmed=x3,k=5)
mudel6=klasterda2(andmed=x3,k=6)

kx2=kmeans(x3,centers=2,iter.max=100,nstart=50)
kx3=kmeans(x3,centers=3,iter.max=100,nstart=50)
kx4=kmeans(x3,centers=4,iter.max=100,nstart=50)
kx5=kmeans(x3,centers=5,iter.max=100,nstart=50)
kx6=kmeans(x3,centers=6,iter.max=100,nstart=50)

gapK=clusGap(x3,kmeans,nstart=50,K.max=6,B=600,spaceH0="original")
print(gapK,method="Tibs2001SEmax") # K=5

# K-keskmiste meetodi tulemuste andmestik
tulemused1=data.frame("Meetod"=rep("K-keskmine",5),"K"=c(2:6))
kx=list(kx2,kx3,kx4,kx5,kx6)
tulemused1$ASW=lapply(X=kx,FUN=function(res)
  summary(silhouette(res$cluster,dist(x3)^2))$avg.width)

# Mudelipõhise klasteranalüüsi tulemuste andmestik
tulemused2=data.frame("Meetod"=rep("Mudelipõhine",5),"K"=c(2:6))
mudelid=list(mudel2,mudel3,mudel4,mudel5,mudel6)
tulemused2$Mudel=lapply(X=mudelid,FUN=function(res) res@model)
tulemused2$Toepara=lapply(X=mudelid,FUN=function(res) res@likelihood)
tulemused2$ICL=lapply(X=mudelid,FUN=function(res) res@criterionValue)
tulemused2$ASW=lapply(X=mudelid,FUN=function(res)
  summary(silhouette(res@partition,dist(x3)^2))$avg.width)

# Joonis 7 ----
joonised(x=x3)
```

Lisa 4. Geenivaramu andmete klasterdamise R-kood

```
library(Rmixmod)

# Funktsioon, mis teeb Rmixmod tulemuste andmestiku
mixmod_tulem3 <- function(mudelid) {
  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
  }
  colnames(tulemused)=c("Mudel","K","Toepara","ICL")
  return(tulemused)
}

# Andmete sisselugemine ja korrastamine ----
load("andmed_nmr.Rdata")
andmed0=cbind(dat1[,c(1:2,4,11,13:14,16)],nmr1[,c(1:4,6,8,10)]); rm(dat1,nmr1)

# Suitsetamise kestuse tunnuse grupeerimine
andmed0$suitskaua[andmed0$kauasuits==0]=1
andmed0$suitskaua[andmed0$kauasuits > 0 & andmed0$kauasuits <= 5]=2
andmed0$suitskaua[andmed0$kauasuits > 5 & andmed0$kauasuits <= 20]=3
andmed0$suitskaua[andmed0$kauasuits > 20]=4

# Kvalitatiivsete tunnuste faktoriteks tegemine
for (i in c(1,3:5,7,15)) {
  andmed0[,i]=as.factor(andmed0[,i])
}

# Puuduvate väärtustega indiviidide eemaldamine ----
andmed=andmed0[complete.cases(andmed0),]; rm(andmed0)
andmed=andmed[andmed$Cit > 0,]
andmed=andmed[andmed$Serum.TG > 0,]

# Biomarkerite ühikute teisendamine ----
andmed$Alb=1000*andmed$Alb # millimooli -> mikromooli liitri kohta
andmed$Cit=1000*andmed$Cit
andmed$Crea=1000*andmed$Crea

# Tabel 4 ja surnute ülevaade ----
apply(andmed[,-c(1,3:5,7,15)],2,FUN=summary)
apply(andmed[,c(1,3:5,7,15)],2,FUN=table)
aggregate(andmed[,c(1,4:5,7)],by=list(andmed$surnud),FUN=table)


# Mudelipõhine klasteranalüüs ----

# Ülesanne 1 ----
yl1=andmed[,c(1:5,7,12:15)]
m1=mixmodCluster(data=yl1[,-3],nbCluster=1:10,
                 dataType="composite",
                 models=mixmodCompositeModel(listModels=c("Heterogeneous_pk_Ekjh_Lk_Bk")),
                 strategy=mixmodStrategy(nbTry=20,nbIterationInAlgo=1000),
                 seed=12,criterion="ICL")
```

```
tul1=mixmod_tulem3(m1["results"])
parim1=m1@bestResult

# Tabel 6 ja 7 ----
table(parim1@partition)
aggregate(yl1[,c(2,7:9)],by=list(parim1@partition),FUN=mean)
aggregate(yl1[,-c(2,7:9)],by=list(parim1@partition),FUN=table)
round(table(parim1@partition,yl1$surnud)[,2]/table(parim1@partition)*100,2)

# Joonis 8 ----
par(mfrow=c(2,2),mar=c(5,5,1,1),cex.lab=2)
boxplot(yl1$vanus~parim1@partition,xlab="Klaster",ylab="Vanus (aastat)")
boxplot(yl1$HDL.C~parim1@partition,xlab="Klaster",ylab="HDL-kolesterool (mmol/l)")
boxplot(yl1$Serum.TG~parim1@partition,xlab="Klaster",ylab="Triglütseriidid (mmol/l)")
boxplot(yl1$Crea~parim1@partition,xlab="Klaster",
        ylab=expression(paste("Kreatiniin (",mu,"mol/l)")))
par(mfrow=c(1,1))


# Ülesanne 2 ----
yl2=andmed[,c(1:5,7:11,15)]
m2=mixmodCluster(data=yl2[,-3],nbCluster=1:10,
                 dataType="composite",
                 models=mixmodCompositeModel(listModels=c("Heterogeneous_pk_Ekjh_Lk_Bk")),
                 strategy=mixmodStrategy(nbTry=20,nbIterationInAlgo=1000),
                 seed=12,criterion="ICL")
tul2=mixmod_tulem3(m2["results"])
parim2=m2@bestResult

# Tabel 8 ja 9 ----
table(parim2@partition)
aggregate(yl2[,c(2,7:10)],by=list(parim2@partition),FUN=mean)
aggregate(yl2[,-c(2,7:10)],by=list(parim2@partition),FUN=table)
round(table(parim2@partition,yl2$surnud)[,2]/table(parim2@partition)*100,2)

# Joonis 9 ----
par(mfrow=c(2,3),mar=c(5,5,1,1),cex.lab=2)
boxplot(yl2$vanus~parim2@partition,xlab="Klaster",ylab="Vanus (aastat)")
boxplot(yl2$Alb~parim2@partition,xlab="Klaster",ylab=expression(paste("Albumiin ( ",mu,"mol/l)")))
boxplot(yl2$Gp~parim2@partition,xlab="Klaster",
        ylab=expression(paste(alpha,"-1 glükoproteiin ( ",mu,"mol/l)")))
boxplot(yl2$Cit~parim2@partition,xlab="Klaster",ylab=expression(paste("Tsitraat (",mu,"mol/l)")))
boxplot(yl2$VLDL.D~parim2@partition,xlab="Klaster",ylab="VLDL-partikli diameeter (nm)")
par(mfrow=c(1,1))

# Tabel 5 ----
table(parim1@partition,parim2@partition)
table(andmed$surnud[parim1@partition==2 & parim2@partition==7])
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Sören Mirski,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Mudelipõhine klasteranalüüs“, 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.

Sören Mirski
15.05.2019