Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Anne Ott

Eestlaste verepildi kirjeldav analüüs

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendaja: Sven Laur

Tartu 2019

Eestlaste verepildi kirjeldav analüüs

Bakalaureusetöö

Anne Ott

Lühikokkuvõte. Tavapäraseim viis patisendi tervisest ülevaatliku pildi saamiseks on teha vereproov. Käesoleva bakalaureusetöö eesmärk on teha eestlaste vereanalüüsidele esmaseid analüüse ja tulemusi visualiseerida. Töös rakendatakse vereanalüüside klasterdamist haige ja terve patsiendi vereks ja kontrollitakse klasterduse korrektsust. Lisaks vaadeldakse lähemalt haigete patsientide verd ja verepildi muutumist ajas. Täpsemalt uuritakse ka angiini ja aneemia diagnoosiga patsiente.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika
Märksõnad: Gaussi segumudel, lineaarne diskriminantanalüüs, peakomponentide analüüs, vereanalüüsid.

Descriptive analysis of Estonian blood tests

Bachelor's thesis

Anne Ott

Abstract. Most common way to determine patient's health status is to do a blood test. The purpose of this bachelor's thesis is to do primary analysis on Estonian blood tests and visualize the results. Clustering blood analysis to healthy and unhealthy patients blood is performed and the validity of results is checked. In addition, unhealthy blood and its changes in time are analyzed further. Patients with diagnosis of tonsillitis and anemia are examined more closely.

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics
Keywords: Gaussian mixture model, linear discriminant analysis, principal component analysis, blood analysis.

Sisukord

Sissejuhatus

Patisendi tervisest ülevaatliku pildi saamiseks alustatakse sageli esmalt vereanalüüsi tegemisest. Vereanalüüsis mõõdetavad erinevate verenäitajate väärtused viitavad, mis haigust patsient võib põdeda. Haigete inimeste veri pole alati ühesugune, üks haigus mõjutab ühte verenäitajat, teine teist näitajat ja kolmanda haiguse puhul võib verepilt täiesti korras olla. Samuti võivad vereanalüütide väärtused haiguse käigus muutuda. Näitajate väärtused võivad erineda sõltuvalt patsiendi soost ja vanusest.

Bakalaureusetöö eesmärk on kirjeldada eestlaste normaalseid ja anormaalseid verenäitajaid ning nende muutumist ajas erinevate diagnooside puhul. Töös tehakse andmetele esmane analüüs, et saada ülevaade andmetest ja tuvastada huvipakkuvaid aspekte. Samuti katsetatakse erinevaid viise andmete visualiseerimiseks. Lähemalt uuritakse angiini ja aneemia diagnoosiga patsientide verepilti.

Töö põhiosa on jaotatud neljaks peatükiks. Esimeses peatükis tutvustatakse töös kasutatavaid andmeid. Teises peatükis antakse teoreetiline ülevaade töös kasutatavatest meetoditest: peakomponentide analüüsist, lineaarsest diskriminantanalüüsist ja lõplikest Gaussi segumudelitest. Kolmandas peatükis tegeletakse vereanalüüside normaalseks ja anormaalseks klasterdamisega. Neljandas peatükis klasterdatakse anormaalset verd ja uuritakse verepildi muutumise trajektoore.

Analüüsi ja jooniste tegemiseks on kasutatud tarkvaraprogrammi R (versioon 3.5.1) ja töö on kirjutatud tekstitöötlusprogrammiga Latex.

Autor soovib tänada töö juhendajat Sven Lauri pühendatud aja, kasulike nõuannete, arvukate selgituste ja suunamise eest.

1 Andmestiku ülevaade

Töös uuritavad andmed pärinevad Tervise ja Heaolu Infosüsteemide Keskusest (TEHIK). Andmete kasutamiseks on saadud eetikakomitee luba „Personaalne meditsiin parema elukvaliteedi tagamiseks ennetava modelleerimise kaudu". Kogu TEHIK-u andmestikus on 97 803 481 rida erinevate analüüsidega (pikas formaadis andmed ehk ühele reale vastab ühe vereanalüüdi väärtus). Nendest 43% moodustavad erinevat sorti vereanalüüsid, millest populaarseimad on hemogramm viieosalise leukogrammiga, kliiniline vere ja glükohemoglobiini analüüs, vere automaatuuring 3-osalise leukogrammiga ja hematoloogilised uuringud.

Andmestikus on tunnusteks

- patsiendi ID,
- diagnoosi kood,
- sugu,
- sünniaasta,
- vereanalüüdi nimi,
- vereanalüüdi väärtus,
- vereanalüüsi tegemise aeg.

Töös uuritavad vereanalüüdid (hemoglobiin, valgevereliblede arv jne) pärinevadki eelmainitud erinevatest vere põhiparameetrite uuringutest. Erinevates vereanalüüside uuringutes mõõdetakse mõnevõrra erinevaid näitajaid. Samuti antakse mitmeid vereanalüüte uuringutes nii protsentides kui ka arvudes. Antud töös on võetud uurimise alla kõik verenäitajad, mis pole mõõdetud protsentides ja mis kuuluvad hemogramm viieosalise leukogrammi alla. Töös uuritavad analüüdid on toodud tabelis 1. Tihti esineb ka teiste uuringute nimede all töös uuritavate analüütide mõõtmisi, ka need on andmestikku sisse võetud. Näiteks sisaldab kliiniline vere ja glükohemoglobiini analüüs kõiki töös vaatluse alla võetud analüüte ja hematoloogilised uuringud on üldisem mõiste vere uuringutest, mille alla kuulub ka hemogramm viieosalise leukogrammiga.

Käesolevas töös kasutatavas andmestikus on kokku 3 115 256 rida patsientide vereanalüüside tulemustega (laias formaadis ehk ühel real on ühe patsiendi kõigi analüütide väärtused), kus on mõõdetud meile huvipakkuvaid analüüte. Esineb ka puuduvaid väärtusi, seega jätame välja analüüdid (veerud) ja patsiendid (read), mille puhul on üle 50% väärtustest puudu ja ülejäänutel puhul asendame need algses lahenduses mediaaniga (puuduvate väärtuste probleemist lähemalt peatükis 4.3). Välja jäävad lümfotsüütide (LYMPH%) ja hemoglobiini (HB) mõõtmised, kuna nende väärtused on rohkem kui pooltel juhtudel

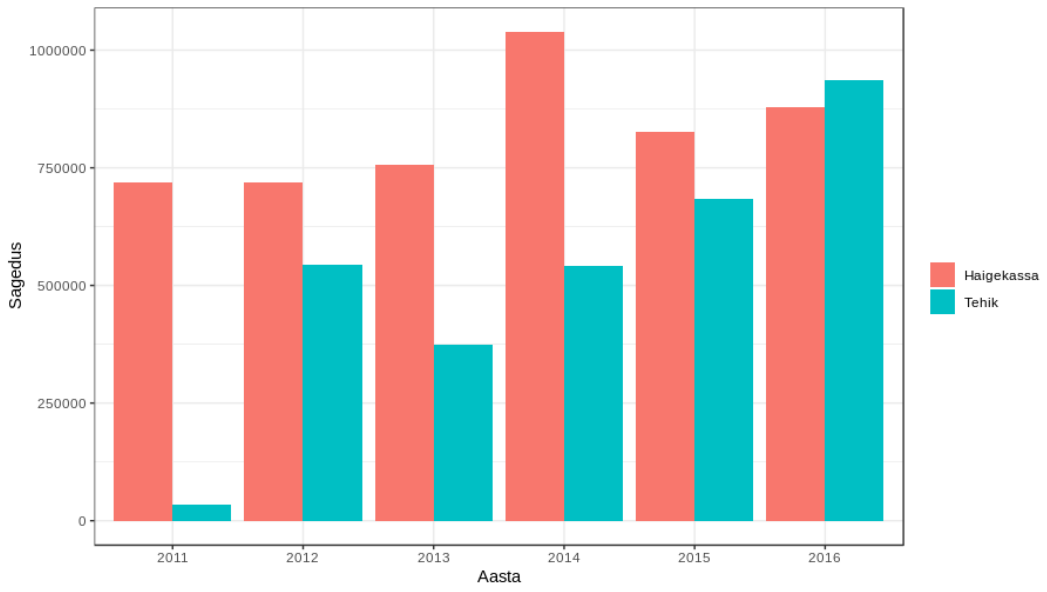
Tabel 1: Vereanalüüdid

Analüüdi lühend	Analüüdi nimi	Ühik
RBC	punaste vereliblede absoluutarv	10^{12}/L
WBC	valgete vereliblede absoluutarv	10^{9}/L
BASO#	basofiilide arv	10^{9}/L
MONO#	monotsüütide absoluutarv	10^{9}/L
EO#	eosinofiilide arv	10^{9}/L
PLT	trombotsüütide arv	10^{9}/L
LYMPH#	lümfotsüütide arv	10^{9}/L
NEUT#	neutrofiilide arv	10^{9}/L
MCH	keskmine hemoglobiini mass ühe punase verelible kohta	pg (pikogramm)
HB	hemoglobiin	g/L
MCHC	keskmine hemoglobiini kontsentratsioon erütrotsüüdis	g/L
MCV	punavereliblede keskmine ruumala	fL (femtoliiter)
MPV	trombotsüütide keskmine maht veres	fL (femtoliiter)

puudu.

Töös kasutatavad TEHIK-u vereanalüüside andmed pärinevad aastatest 2005-2016. Aastatest 2005-2010 on väga vähe mõõtmisi, alla 2 000 vereanalüüsi aasta kohta. Enne aastat 2012 ongi TEHIK-u andmed puudulikud, kuid alates 2012 aastast on seal kajastatavate epikriiside hulk märgatavalt kasvanud [1]. Kuna andmed enne 2012 pole juhuslikud, siis tegeleme töös andmete valideerimisega.

Tehtud analüüside andmed on kajastatud ka Eesti Haigekassa raviarvetes. Nimelt peab tervishoiuteenuse osutaja Haigekassale esitama raviarve, et saada Haigekassalt teenuse eest hüvitist. Haigekassa raviarvete andmed on mõeldud raamatupidamiseks ja seega on üldjuhul täpsemad. Joonisel 1 on võrreldud aastate 2011-2016 TEHIK-u vereanalüüside andmete täielikkust Haigekassa vereanalüüside raviarvetega (Haigekassa andmed pärinevad Sven Laurilt). Kõikidel aastatel (v.a 2016) on Haigekassa raviarvetes kajastatud rohkem vereanalüüse kui TEHIK-u andmetes. Tuleb märkida, et joonisel on Haigekassa raviarvetes toodud vaid kõige populaarsemate vereuuringud. Kuna uuritavas TEHIK-u andmestikus võivad vereanalüüdid olla ka teiste mittetriviaalsete vereuuringute sees, siis ei ole TEHIK-u sagedused joonisel nii täpsed. Siiski võib jooniselt järeldada, et aastaks 2016 on TEHIK-us andmeid piisavalt, et võiks öelda, et tegemist on kõikse koguga. Siiski pole välistatud, et Haigekassa ja TEHIK-u andmetes võib esineda nihkeid.



Joonis 1: Aastas tehtud vereanalüüside arv

Igale diagnoosile vastab RHK-10 kood (*International statistical classification of diseases and related health problems* ehk ICD-10), mis on rahvusvaheline numbri ja tähekombinatsioonist koosnev kood klassifitseerimaks haigusi, vigastusi ja terviseprobleeme [2]. Tänu RHK-10 koodile saab tekstilised diagnoosid muuta tähe-numbri kombinatsioonideks, lihtsustades andmete hilisemat analüüsi. Kõik koodid ja nende tähendused on toodud Eesti Sotsiaalministeeriumi kodulehel [3].

2 Kasutatav metoodika

Käesolevas peatükis antakse teoreetiline ülevaade töös kasutavatest meetoditest. Täpsemalt tutvustatakse peakomponentide analüüsi, lineaarset diskriminantanalüüsi ja normaaljaotuste segumudelit ehk Gaussi segumudelit. Viimast kasutame normaalsete ja anormaalsete vereanalüüside defineerimiseks.

2.1 Peakomponentide analüüs

Paljude andmestike puhul on probleemiks vaadeldavate tunnuste paljusus. Peakomponentide analüüsi (PCA) abil on võimalik andmete dimensiooni vähendada. PCA teeb andmetele lineaarprojektsioon ehk andmed projekteeritakse ortogonaalselt uuele madalamale tasandile. Seejuures katab esimene peakomponent ära kõige suurema tunnuste varieeruvuse, teine komponent suuruselt teise varieeruvuse, olles samas ortogonaalne esimese komponendi suhtes jne. Peakomponentide analüüsi ideega tuli välja Karl Pearson aastal 1901, meetodi formuleeris 1933. aastal Harold Hotelling. Peakomponentide analüüs leidis laiemat rakendust arvutite kasutusele tulekuga.

Käesolev peatükk põhineb Tartu Ülikooli matemaatika ja statistika instituudi dotsendi Imbi Traadi mitmemõõtmelise analüüsi loengukonspektil [4] ja T. W. Andersoni raamatul *„An Introduction to Multivariate Statistical Analysis"* [5], kui pole viidatud teisti.

Eeldame, et meil on tegu sõltuvate tunnustega. Olgu meil m elemendiga juhuslik vektor $\boldsymbol{X} = (X_1, ..., X_m)^T$. Kuna peakomponentide analüüsi puhul pakuvad huvi ainult dispersioon ja kovariatsioon, siis üldisust kitsendamata võime eeldada, et andmed on tsentreeritud ehk

$$E\boldsymbol{X} = 0 \ .$$

Tunnuse X dispersiooni- ehk kovariatsioonimaatriks avaldub

$$D\boldsymbol{X} = E\big((\boldsymbol{X} - E\boldsymbol{X})(\boldsymbol{X} - E\boldsymbol{X})^T\big) = E(\boldsymbol{X}\boldsymbol{X}^T) = \boldsymbol{\Sigma} \ ,$$

kus peadiagonaalil on dispersioonid

$$DX_i = \sigma_{ii}$$

ja väljaspool kovariatsioonid

$$\operatorname{cov}(X_i, X_j) = \sigma_{ij} \ .$$

Kovariatsioonimaatriksi abil saame mõõta tunnuste vahelist sõltuvust.

Peakomponendid *(principal components)* $P_2, P_2, ..., P_m$ on omavahel mittekorreleeritud uued tunnused, mis on esialgsete tunnuste X_i lineaarkombinatsioonid. Lineaarkombinatsioonid on kujul $P = \boldsymbol{\alpha}^T \boldsymbol{X}$, kus kordajad $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)^T$ on valitud nii, et P_1 dispersioon oleks maksimaalne, P_2 dispersioon suuruselt järgmine jne. Seejuures peavad kordajad $\boldsymbol{\alpha}$ olema normeeritud ehk $\boldsymbol{\alpha}^T \boldsymbol{\alpha} = 1$.

Olgu dispersioonimaatriksi $\boldsymbol{\Sigma}$ omaväärtused $\lambda_1 \geq \lambda_2, ..., \geq \lambda_m$ ja omaväärtustele vastavad ortogonaalsed omavektorid $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_m$. Järgnevad tõestused pärinevad Joliffe raamatust [6].

Teoreem. Esimese komponendi P_1 maksimaalne dispersioon on dispersioonimaatriksi $\boldsymbol{\Sigma}$ suurim omaväärtus $DP_1 = \lambda_1$.

Tõestus. Vastavalt peakomponentide definitsioonile, peame me maksimeerima DP_1 tingimusel $\boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 = 1$. Dispersioon on avaldatav kujul $DP_1 = D(\boldsymbol{\alpha}_1^T \boldsymbol{X}) = D(\boldsymbol{\alpha_1}^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_1)$. Kasutades Lagrange'i meetodit maksimeerime funktsiooni

$$\boldsymbol{\alpha_1}^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 - 1) \ , \tag{1}$$

kus λ on Lagrange'i kordaja. Leiame avaldise (1) tuletise $\boldsymbol{\alpha}_1$ suhtes ja võrdsustame nulliga

$$\boldsymbol{\Sigma} \boldsymbol{\alpha}_1 - \lambda \boldsymbol{\alpha}_1 = 0 \ .$$

Järelikult on otsitavaks lahendiks $\boldsymbol{\Sigma}$ omavektor $\boldsymbol{\alpha}_1$ ja Lagrange'i kordajaks omaväärtus λ. Tahame leida omaväärtuse, mille puhul $P_1 = \boldsymbol{\alpha}_1^T \boldsymbol{X}$ dispersioon on maksimaalne. Selleks maksimeerime $\boldsymbol{\alpha}_1^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_1$ ja saame

$$\boldsymbol{\alpha}_1^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1^T \lambda \boldsymbol{\alpha}_1 = \lambda \boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 = \lambda$$

ehk λ peab olema suurim $\lambda_1, \lambda_2 ..., \lambda_m$. Seega $\boldsymbol{\alpha}_1$ on omavektor suurimale $\boldsymbol{\Sigma}$ omaväärtusele λ_1 ja $DP_1 = D(\boldsymbol{\alpha}_1^T X) = \boldsymbol{\alpha}_1^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 = \lambda_1$. □

Teoreem. Teise komponendi P_2 maksimaalne dispersioon on dispersioonimaatriksi $\boldsymbol{\Sigma}$ suuruselt teine omaväärtus $DP_2 = \lambda_2$

Tõestus. Teine peakomponent $P_2 = \boldsymbol{\alpha}_2^T \boldsymbol{X}$ peab maksimeerima $\boldsymbol{\alpha}_2^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_2$, kusjuures esimene ja teine peakomponent peavad olema omavahel mittekorreleeritud ehk

$$\text{cov}(\boldsymbol{\alpha}_1^T \boldsymbol{X}, \boldsymbol{\alpha}_2^T X) = 0 \ .$$

Samas teame, et

$$\text{cov}(\boldsymbol{\alpha}_1^T\boldsymbol{X}, \boldsymbol{\alpha}_2^T\boldsymbol{X}) = \boldsymbol{\alpha}_1^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2^T\lambda_1\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_2^T\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_2 \ ,$$

nullist erineva λ jaoks. Järelikult peavad kehtima

$$\boldsymbol{\alpha}_1^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = 0, \quad \boldsymbol{\alpha}_2^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = 0, \quad \boldsymbol{\alpha}_2^T\boldsymbol{\alpha}_1 = 0, \quad \boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_2 = 0 \tag{2}$$

et P_1 ja P_2 poleks omavahel korreleeritud. Sarnaselt valemile (1) on teise peakomponendi puhul vaja maksimeerida

$$DP_2 = D(\boldsymbol{\alpha}_2^T\boldsymbol{X}) = \boldsymbol{\alpha}_2^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda(\boldsymbol{\alpha}_2^T\boldsymbol{\alpha}_2 - 1) - \phi\boldsymbol{\alpha}_2^T\boldsymbol{\alpha}_1$$

kus λ, ϕ on Lagrange'i kordajad. Leiame eelnevast tuletise $\boldsymbol{\alpha_2}$ suhtes ja võrdsustame nulliga

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1 = 0 \ .$$

Seejärel korrutame vasakult läbi $\boldsymbol{\alpha}_1^T$ ja saame

$$\boldsymbol{\alpha}_1^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_1 = 0 \ . \tag{3}$$

Paneme tähele, et valemite (2) tõttu on esimesed kaks liiget võrdsed nulliga ja seega taandub valem (3) kujule $\phi = 0$. Nagu esimese komponendi puhul, saame ka nüüd, et

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda\boldsymbol{\alpha}_2 = 0 \ ,$$

kus λ on $\boldsymbol{\Sigma}$ omaväärtus ja $\boldsymbol{\alpha}_2$ omaväärtusele vastav omavektor.

Jällegi saame, et $\lambda = \boldsymbol{\alpha}_2^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_2$, kusjuures λ peab olema suurim võimalik. Kuna λ ei saa võrduda λ_1, sest sellisel juhul oleks komponendid korreleeritud, siis järelikult on λ suuruselt teine $\boldsymbol{\Sigma}$ omaväärtus ehk λ_2. Järelikult $DP_2 = \lambda_2$ □

Kovariatsiooni maatriksi $\boldsymbol{\Sigma}$ jälg on lähtetunnuste kogudispersioon $DX_1 + ... + DX_m = \sigma_{11} + ... + \sigma_{mm} = tr(\boldsymbol{\Sigma})$. Lineaaralgebrast teada olevate tulemuste põhjal saame maatriksi $\boldsymbol{\Sigma}$ jälje avaldada omaväärtuste summana $DX_1 + ... + DX_m = \lambda_1 + \cdots + \lambda_m$.

Seega saab lähteandmete dispersiooni DX avaldada omaväärtuste ehk peakomponentide dispersiooni kaudu, kusjuures iga järgmise peakomponendi dispersioon on maksimaalne võimalik.

Eelneva põhjal oleme näidanud, et valem i-nda peakomponendi leidmiseks on

$$P_i = \boldsymbol{\alpha}_i^T \boldsymbol{X}$$

dispersiooniga

$$DP_i = \boldsymbol{\alpha}_i^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 = \lambda_i \ .$$

Kui suure osa koguvarieeruvusest katab ära i-s komponent, väljendab kordaja

$$V_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_m} \ .$$

2.2 Lineaarne diskriminantanalüüs

Lineaarse diskriminantanalüüsi (LDA) rajajaks oli 1936. aastal Sir Ronal Fisher. LDA eesmärgiks on vähendada andmete dimensionaalsust eemaldades üleliigsed ja sõltuvad tunnused. LDA käigus projekteeritakse andmed madalamasse dimensiooni nii, et klassidevaheline eristatavus oleks võimalikult suur. Käesolev peatükk põhineb A. C. Rencheri ja W. F. Christenseni raamatul „*Methods of Multivariate analysis*" [7].

Olgu meil vaatlused jaotatud K klassi vahel nii, et igas klassis on $n_1, ..., n_K$ vaatlust ja olgu $n = n_1 + ... + n_K$ vaatluste koguarv. Esimesse klassi kuuluvad vaatlused $\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, ..., \boldsymbol{x}_{1n_1}$, teise $\boldsymbol{x}_{21}, \boldsymbol{x}_{22}, ..., \boldsymbol{x}_{2n_2}$ jne. Tähistame üldise klasside keskmise, kus

$$\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}$$

ja iga klassi keskmise $\bar{\boldsymbol{x}}_1, ..., \bar{\boldsymbol{x}}_K$, kus

$$\bar{\boldsymbol{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}, \ \ i = 1, ..., K \ .$$

Olgu meil K klassi koguhajuvus maatriks $\boldsymbol{S}$ (*total sum of squares*), siis

$$\boldsymbol{S} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}})^T \ .$$

Keskmise algebralistest omadustest lähtuvalt saab seda esitada

$$\boldsymbol{S} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^T + \sum_{i=1}^{K} n_i (\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})^T \ . \tag{4}$$

Esimene liidetav valemis (4) on klassidesisene hajuvus

$$\boldsymbol{W} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^T \ ,$$

mis mõõdab klassi elementide ruutkeskmisi kaugusi klassi keskpunktist $\bar{\boldsymbol{x}}_i$.

Teine liidetav valemis (4) on klassidevaheline hajuvus

$$\boldsymbol{B} = \sum_{i=1}^{K} n_i (\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})^T \ ,$$

mis mõõdab klassi keskpunktide ruutkeskmisi erinevusi andmete keskpunktist.

Eesmärgiks on leida selline tunnuste lineaarne kombinatsioon $\boldsymbol{\alpha}^T \boldsymbol{x}$, mille korral oleks klassidevahelise hajuvuse $\boldsymbol{B}$ ja klassidesisese hajuvuse $\boldsymbol{W}$ suhe maksimaalne iga projekteeritud vaatluse z_{ij} jaoks, mis avaldub

$$z_{ij} = \boldsymbol{\alpha}^T \boldsymbol{x}_{ij} \ .$$

Analoogselt avalduvad keskmised $\bar{z}_i = \boldsymbol{\alpha}^T \bar{\boldsymbol{x}}_i$. Seega lähtuvalt eesmärgist tahame maksimeerida suhet

$$\lambda = \frac{\boldsymbol{\alpha}^T \boldsymbol{B} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \boldsymbol{W} \boldsymbol{\alpha}} \ . \tag{5}$$

Lihtsustades eelenvat saame

$$\boldsymbol{\alpha}^T (\boldsymbol{B}\boldsymbol{\alpha} - \lambda \boldsymbol{W}\boldsymbol{\alpha}) = \boldsymbol{0} \ .$$

Kuna lahend $\boldsymbol{\alpha}^T = \boldsymbol{0}$ pole lubatud, siis kõik lahendid on leitavad vastavalt

$$(\boldsymbol{W}^{-1}\boldsymbol{B} - \lambda \boldsymbol{I})\boldsymbol{\alpha} = \boldsymbol{0} \ .$$

Eelneva võrrandi lahendid on $\boldsymbol{W}^{-1}\boldsymbol{B}$ omaväärtused. Seega avaldist (5) maksimeerib suurim omaväärtus λ_1 ja sellele vastav omavektor $\boldsymbol{\alpha}_1$. Järgnev projektsioon $\boldsymbol{\alpha}^T \boldsymbol{x}$ peab olema mittekorreleeritud eelenva projektsiooniga $\boldsymbol{\alpha}_1^T \boldsymbol{x}$ ja jälle maksimeerima suhet avaldises (5). Selleks vektoriks sobib suuruselt teisele omaväärtusele λ_2 vastav omavektor $\boldsymbol{\alpha}_2$.

Võimalike sisukate omaväärtuste arvu määrab maatriksi $\boldsymbol{B}$ astak s, mis piirab ära nullist erinevate omaväärtuste $\lambda_1, ..., \lambda_s$ arvu. Algebralistest omadustes tuleneb, et astaku s väärtus peab olema väiksem võrdne $K-1$ ja tunnuste arvust p. Seega on lahenditeks maatriksi $\boldsymbol{W}^{-1}\boldsymbol{B}$ omaväärtused $\lambda_1, ..., \lambda_s$ ja omavektorid $\boldsymbol{\alpha_1}, ..., \boldsymbol{\alpha_s}$. Olemegi saanud s mittekorreleeritud lineaarset diskriminant funktsiooni $z_1 = \boldsymbol{\alpha}_1^T \boldsymbol{x}, ..., z_s = \boldsymbol{\alpha}_s^T \boldsymbol{x}$.

Iga diskriminant funktsiooni z_i, $i = 1, ..., s$, tähtsust väljendab V_i

$$V_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_s} \ .$$

2.3 Lõplikud segumudelid

Vereproovide andmetes on segamini kaks komponenti – haigete ja tervete inimeste vereanalüüside tulemused. Tegelikkuses on andmetes puudu informatsioon, kas patsient oli vereproovi tegemise hetkel haige või mitte. Andmete klasterdamisel soovime iga patsiendi mõõtmise puhul kindlaks teha, kumma segumudeli komponendi alla inimene kuulub, selleks kasutame mürakomponendiga laiendatud Gaussi segumudelit.

2.3.1 Gaussi segumudel koos müra komponendiga

Olgu Z juhuslik suurus, mis võtab väärtusi $1, ..., K$ tõenäosusega $P(Z = k) = \pi_k$, $k = 1, ..., K$. Olgu meil juhuslik vektor $\boldsymbol{X}$, mis on segu K komponendist $\boldsymbol{X}_1, ..., \boldsymbol{X}_K$ tihedusfunktsioonidega $f_1, ..., f_K$. Kõigepealt defineerime üldise lõpliku segumudeli. Vektori $\boldsymbol{X}$ jaotus on kirjeldatav lõpliku K-komponendilise segumudeli tihendusfunktsiooni kaudu

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

kus $\boldsymbol{\theta}_k$ on tihedusfunktsiooni f_k parameetrite vektor ja $\pi_1, ..., \pi_K$, $0 \leq \pi_k \leq 1$, on segu kaalud ja

$$\sum_{k=1}^{K} \pi_k = 1 \ .$$

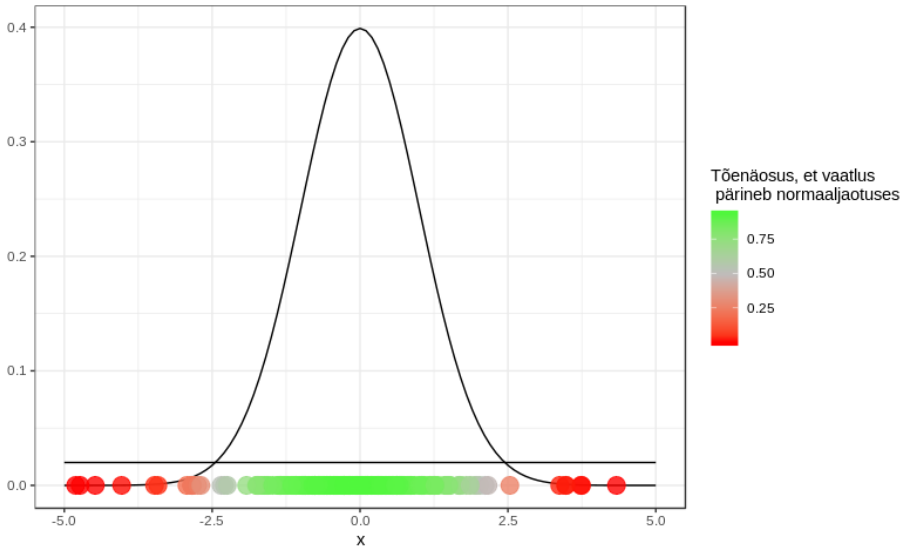
Gaussi segumudeli korral valitakse komponendid mitmemõõtmelisest normaaljaotusest $\boldsymbol{X}_k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, kus k-nda komponendi tihedusfunktsioon on

$$f_k(\boldsymbol{x}|\boldsymbol{\theta}_k) = f_k(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\det(2\pi\boldsymbol{\Sigma}_k)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right) \ ,$$

kus $\boldsymbol{\mu}_k$ on k-nda komponendi keskväärtuste vektor ja $\boldsymbol{\Sigma}_k$ on k-nda komponendi kovariatsioonimaatriks [9].

Tihti kasutatakse Gaussi segumudeli robustsemaks muutmiseks ühtlase jaotusega komponenti ehk mürakomponenti. Mürakomponendiks f_0 on ühtlane jaotus tihedusparameetriga p, see tähendab $f_0(x) = p$ iga vaatluse korral. Teoreetiliselt saab mürakompo-

nendi korral tihedust p määrata hinnates andmeid sisaldava kujundi (hüperkuubi või kera) mahtu. Praktikas võib p väärtuse fikseerida suvaliselt. Väikse p väärtuse korral lähevad müra klastrisse vaatlused, mis on väga erinevad kõikidest Gaussi komponentidest. Suurema p väärtuse korral jagatakse Gaussi komponentide vahel vaid neid punkte, mis on klastri keskmistele väga lähedal. Oluline on tähele panna, et erineva variatsiooniga klastrite korral on mürakomponendi mõju erinev, kuna jaotus toimub tihedusfunktsiooni väärtuse järgi. Tegu on kunstliku jaotusega, mis formaalselt täidab segumudeli nõudeid, aga mille ainus mõte on koguda erindid ühte kokku. Näide mürakomponendist on toodud joonisel 2, kus p väärtuseks on valitud 0.02. Värvilised punktid on genereeritud nii normaaljaotusest kui ka ühtlasest jaotusest. Seejärel on Gaussi segumudeli abil leitud tõenäosused, et punkt pärineb normaaljaotusest. Punaste punktide on väärtused, mis määratakse mürakomponendi alla, sest tõenäosus pärineda normaaljaotusest on väike.



Joonis 2: Gaussi segumudel koos mürakomponendiga

2.3.2 Parameetrite hindamine

Gaussi segumudeli parameetri $\boldsymbol{\theta} = (\pi_1, \pi_2 ..., \pi_K, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, ..., \boldsymbol{\Sigma}_K)$ hindamiseks sobib suurima tõepära hinnang. Kuna logaritmilise tõepära maksimeerimine on ekvivalentne tavalise tõepära maksimeerimisega, siis kasutame edaspidi just logaritmilist tõepära, mis avaldub kujul

$$\ln \mathscr{L}(\boldsymbol{\theta} | x_1, ..., x_n) = \sum_{i=1}^{n} \left(\ln \left(\sum_{k=1}^{K} \pi_k f_k(x_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \right) .$$

Selle otsene maksimeerimine on aga tihti numbriliselt keeruline. Iga vaatluste $x_1, ..., x_n$ jaoks on olemas vastav teadmata klass ehk latentsed tunnused $z_1, ..., z_n$, mis näitavad, kust

vaatlus tegelikult pärineb.

Suurima tõepära hinnangu leidmiseks kasutatakse EM-algoritmi. EM-algoritmi iteratsioon on järgnev:

1. Olgu algne parameetri väärtus $\boldsymbol{\theta}$.
2. E-Samm: Kasutades käesolevaid parameetrite väärtusi leitakse järeltõenäosused. Järeltõenäosus, et vaatlus $\boldsymbol{x}_i$ kuulub klassi $k = 1,...,K$ on tinglik tõenäosus $P(z_i = k|\boldsymbol{x}_i,\boldsymbol{\theta})$, mis avaldub järgnevalt

$$P(z_i = k|\boldsymbol{x}_i,\boldsymbol{\theta}) = \frac{\pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} = \gamma_{ik}$$

3. M-Samm: Kasutades eelmises sammus leitud järeltõenäosuseid hinnatakse keskväärtused, dispersioonid ja segu kaalud ümber vastavalt

$$\begin{aligned}
n_k &= \sum_{i=1}^{n} \gamma_{ik} \ , \\
\boldsymbol{\mu}_k^{'} &= \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{ik} \boldsymbol{x}_i \ , \\
\boldsymbol{\Sigma}_k^{'} &= \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{'})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{'})^T \ , \\
\pi_k^{'} &= \frac{n_k}{n} \ .
\end{aligned}$$

EM-sammude iteratsiooni korratakse kuni saavutatakse peatumistingimus. Selleks tingimuseks on tavaliselt kas tõepära stabiliseerumine või parameetrite väärtuste stabiliseerumine. Paljudel juhtudel tuletab algoritm mõistliku parameetri hinnangu ja kokkuvõtteks on see laialt levinud meetod keskmistatud suurima tõepära hindamiseks. [8, 9]

2.3.3 Bayesi informatsioonikriteerium

Bayesi informatsioonikriteerium (BIC) kasutatakse siis, kui mudelit on vaja sobitada logaritmitud suurima tõepära meetodiga. Kui klastrite arv on teadmata, siis saab BIC abil selle leida. Mudeli parameetriga $\boldsymbol{\theta}$ korral avaldub Bayesi informatsioonikriteerium

$$\mathrm{BIC} = 2\ln\left(\mathscr{L}\left(\hat{\boldsymbol{\theta}}|\boldsymbol{x}\right)\right) - m\log(n) \ ,$$

kus $\mathscr{L}(\hat{\boldsymbol{\theta}}|\boldsymbol{x})$ on suurima tõepära hinnang, m on vabade parameetrite arv mudelis ja n on vaatluste arv.

BIC abiga saame leida mudeli tüübi, mõjusa hinnangu komponentide arvule K (kui see on teadmata) ja kovariatsiooni hinnangu. BIC karistab keerulisemaid mudeleid ja eelistab valikus olevaid lihtsamaid mudeleid [10]. Sobivaima klastrite arvu leidmisel arvutatakse EM-algoritmi käigus välja BIC eraldi kõigi erinevate klastrite arvude $2,3,...,M$ ja klastri optimaalsete parameetrite korral, kus M on eelnevalt valitud maksimaalne soovitud klastrite arv. Parimaks mudeliks valitakse maksimaalse BIC väärtusega mudel [11].

2.4 Standardiseerimine

Vereanalüütide referentsvahemikud on erinevate laiustega ja erinevates suurusjärkudes. Eri suuruseid saab võrreldavaks muuta mitmel viisil, mis rõhutavad erinevaid aspekte. Klassikaline standardiseerimine

$$\text{standardiseeritud väärtus} = \frac{\text{tegelik väärtus} - \text{keskväärtus}}{\text{standardhälve}} \tag{6}$$

kirjeldab väärtuste hajuvust ja relatiivseid muutusi.

Keemiliste suuruste korral huvitavad meid aga ka absoluutsed muutused. Sellisel juhul on vaja teada referentsväärtust ehk väärtust, mis peaks olema mõõdetava suuruse kõige tüüpilisem väärtus. Andmete standardiseerimine valemi

$$\text{standardiseeritud väärtus} = \frac{\text{tegelik väärtus} - \text{referentsväärtus}}{\text{referentsväärtus}} * 100 \ .$$

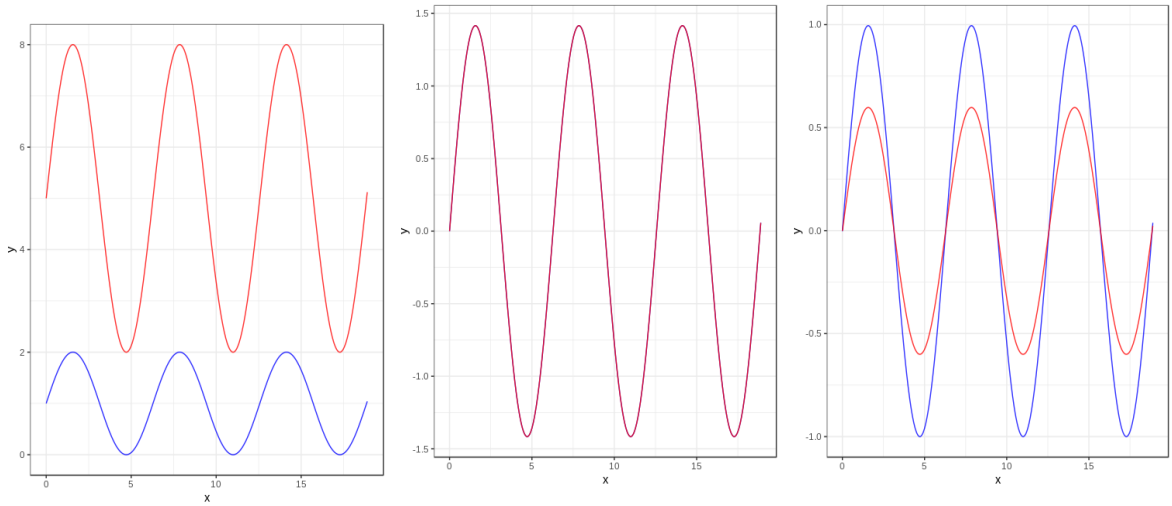
järgi jätab alles informatsioon väärtuste absoluutse muutumise kohta.

Referentsväärtuseks võiks valida keskväärtuse, kuid see on mõjutatav erinditest, mida anormaalses veres palju esineb. Seega kasutame alternatiivses standardiseerimises keskväärtuse asemel mediaani

$$\text{standardiseeritud väärtus} = \frac{\text{tegelik väärtus} - \text{normaalse vere mediaan}}{\text{normaalse vere mediaan}} * 100 \ , \tag{7}$$

kusjuures normaaljaotuse korral on keskväärtus võrdne mediaaniga.

Erinevusest aitab paremini aru saada joonisel 3 toodud näide. Klassikaline standardiseerimise puhul langevad standardiseeritud väärtused kokku (sellest ka lilla värvus). See ei võtta aga arvesse asjaolu, et punase joone absoluutne varieeruvus on tunduvalt suurem. Mediaani järgi standardiseerimisel ilmneb absoluutne varieeruvus.



(a) Algsed andmed

(b) Klassikaline standardiseerimine

(c) Mediaani järgi standardiseerimine

Joonis 3: Standardiseerimine

3 Klasterdamine normaalseks ja anormaalseks vereks

Vereanalüüside puhul on teadmata, kas veri on mõõdetud tervelt või haigelt inimeselt. Eraldamaks tervete ja haigete inimeste vereanalüüside tulemusi on vaja analüüside jaotused eraldada. Bioloogilised mõõtmised jälgivad enamasti normaaljaotust, ja seega kasutame veretüüpide kaheks jaotamiseks Gaussi segumudelit. Katsetati ka alternatiivseid mudeleid, kuid kuna suuri erinevusi Gaussi segumudeliga ei ilmnenud, otsustasime jääda levinuima mudeli juurde. Anormaalsed verenäitajad on erandlike väärtustega näitajad, seega lisame mudelile ka mürakomponendi (ühtlasest jaotusest väärtused). Mürakomponendi abil saab visata välja erindeid, mis muidu kõigutaks oluliselt punkthinnanguid. Tulemuseks saadud parameetrite hinnangud on meie mudeli puhul robustsemad.

Gaussi segumudeli abil klasterdamiseks kasutasime R-i paketti *mclust*, kus mudeli parameetrite hindamiseks kasutatakse EM-algoritmi. Iga klastri geomeetrilised tunnused (maht, kuju, orientatsioon) määratakse kovariatsioonimaatriksi $\mathbf{\Sigma}_k$ abil, $k = 1, ..., K$. Kovariatsioonimaatriksi $\mathbf{\Sigma}_k$ omaväärtuste dekompositsiooni abil saame kirjutada

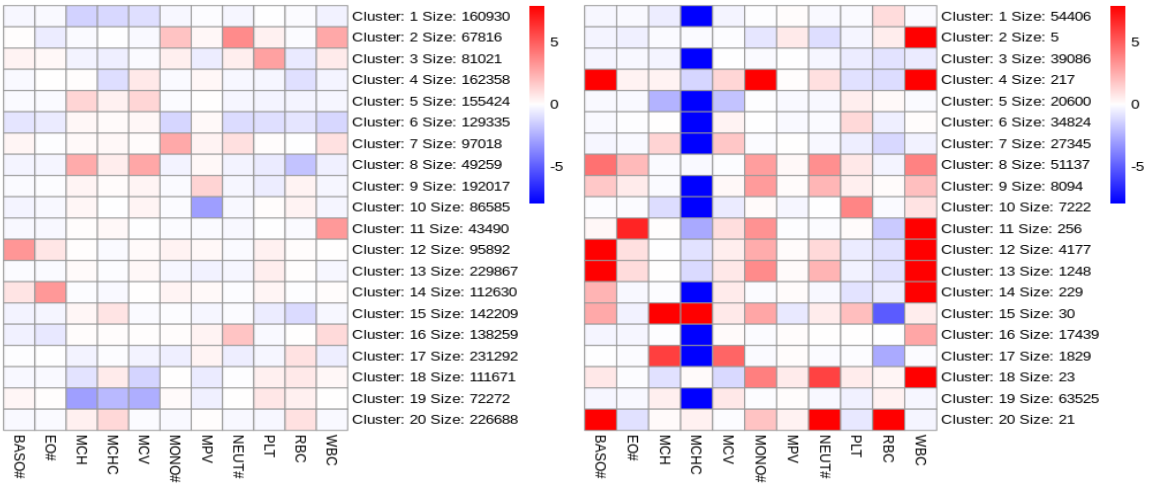
$$\mathbf{\Sigma}_k = \lambda_k, \boldsymbol{D}_k \boldsymbol{A}_k \boldsymbol{D}_k^T \ ,$$

kus $\boldsymbol{D}_k$ on ortogonaalne omavektorite maatriks, $\boldsymbol{A}_k$ diagonaalmaatriks, mille peadiagonaali elementideks on omaväärtused ja λ_k on komponentide konstantne osakaal . Paketis olevad mudelid on järgnevad: EII, VII, EEI, VEI, EVI, VVI, EEE, EEV, VEV ja VVV. Tähekombinatsiooni esimene täht viitab mahule, teine täht kujule ja kolmas täht orientatsioonile. E (*equal*) tähistab võrdsust, V (*variable*) tähistab muutlikust ja I (*coordinate axes*) koordi-

naatteljestikku. Kovariatsioonimaatriksit kolm parameetrit on vasatavalt – λ_k ehk maht, $\boldsymbol{A}_k$ ehk kuju ja $\boldsymbol{D}_k$ ehk orientatsioon. Pakett kasutab suurima tõepära meetodit, et sobitada andmetele kõiki mudeleid. Optimaalne mudel valitakse suurima BIC-i järgi. [8]

Klasterdamise tulemusena määrati normaalseks 2 585 319 analüüsi ja anormaalseks vereks 332 426 analüüsi. Seega peab meie algoritm andmestikust 11.4% veremõõtmisi anormaalseks.

Normaalse ja anormaalse vere analüütide visualiseerimiseks kasutame soojuskaarti (*heatmap*). Analüütide väärtused on erinevates suurusjärkudes, seega standardiseerime eelnevalt kõigi analüütide väärtused klassikalise standardiseerimise (valem 6) järgi, sest huvi pakub just analüütide hajuvus. Andmete paljususe tõttu pole võimalik kõiki andmestiku ridu soojuskaardil kujutada. Seetõttu kasutame soojuskaardi tegemisel R-i funktsiooni *pheatmap* argumenti *kmeans_k*, mis andmestiku sarnased read üheks klastriks kokku võtab ja joonisel klastri keskmisi kujutab. Soojuskaardilt 4 paistavad anormaalse ja normaalse vere analüütide üldised mustrid. Normaalse vere väärtused on vähem varieeruvad võrreldes anormaalse vere väärtustega. Lisaks iseloomustab anormaalset verd see, et vähemalt ühe analüüdi väärtused on paigast ära. Tulemus on oodatav, sest haigestumise puhul ei lähe kõik vereanalüüdid paigast ära vaid ainult teatud. Jooniseid vaadates tuleb tähele panna, et *kmeans* algoritm ei tagasta alati ühesuuruseid klastreid, kuid joonisel on kõik klastrid kujutatud samas suuruses ridadena. Seetõttu on mõned mustrid visuaalselt ala ja mõned üle esindatud. Seega näitab antud joonis vaid tüüpilisi vaatluste koosesinemiste mustreid ja mitte nende esinemissagedust.



(a) Normaalne veri

(b) Anormaalne veri

Joonis 4: Vereanalüütide väärtused

3.1 Referentsvahemikud

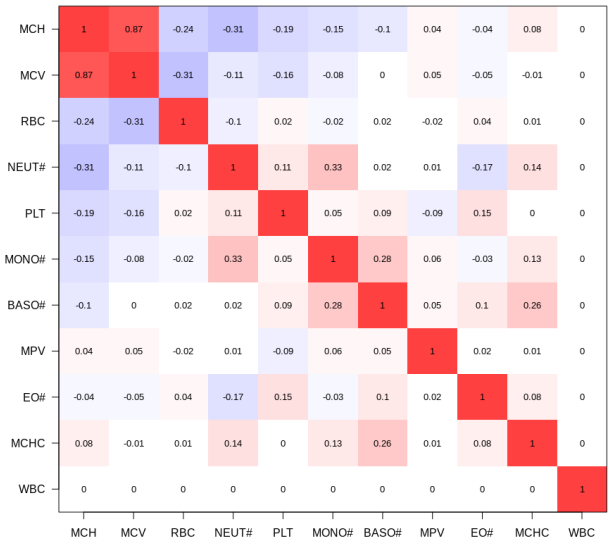
Meditsiinis kasutatakse analüütide tõlgendamiseks referentsvahemikke. Referentsvahemik on vahemik, mis näitab vastava vereanalüüdi oodatavaid väärtusi. Väärtuse jäämine referentsvahemikest välja viitab kõrvalekaldele normaalsest verest ja võib olla indikaatoriks erinevatele haigustele.

Leiame mudeli poolt normaalseks määratud vere analüütide referentsvahemikud, kasutades valemit

$$U = \bar{X} \pm \sigma \ , \tag{8}$$

kus $\bar{X}$ on normaalse vere analüüdi keskväärtus ja σ standardhälve. Mudeli leitud referentsvahemike võrdlemiseks tegelike referentsvahemikkutega kasutame Tartu Ülikooli kliinikumi (TÜK) poolt määratud referentsvahemikke [12]. Referentsvahemike defineerimise eesmärgiks on kontrollida, kas normaalse vere komponent sobitub TÜK-i raamidesse ja ega andmetes pole selget nihet.

Usaldusintervalli leidmine valemi 8 põhjal jätab arvestamata tunnuste omavahelise korrelatsiooni. Joonisel 5 on toodud välja analüütide korrelatsioonimaatriks. Kuna maatriksi peadiagonaalil on ühed ja väljaspool peadiagonaali on väärtused -0.3 ja 0.3 vahel, siis on vereanalüüdid omavahel nõrgalt või üldse mitte korreleeritud. Erandiks on vaid punaverelible keskmine ruumala (MCV) ja keskmine hemoglobiini mass punase verelible kohta (MCH), mille vahel on tugev korrelatsioon 0.87. Sõltuvust põhjendab asjaolu, et mida suurem on punane verelible, seda rohkem hemoglobiini ta tavaliselt ka sisaldab.



Joonis 5: Normaalse vere analüütide korrelatsioonimaatriks

Joonisel 6 on võrreldud mudeli ja kliinikumi referentsvahemike laiuste kattuvust. Kuna analüütide referentsvahemike suurusjärgud on erinevad (näiteks BASO# 0-0.1, MCHC aga 317-357), normeerime kõik väärtused nii, et TÜK-i referentvahemiku laius oleks joonisel 2 ühikut. TÜK-i referentsvahemike alumisele piirile vastab -1, keskmisele 0 ja ülemisele piirile 1. Ideaalsel juhul peaksid mudeli referentsvahemikud (värvilised jooned) olema keskel ja mitte sealt suurel määral välja minema.

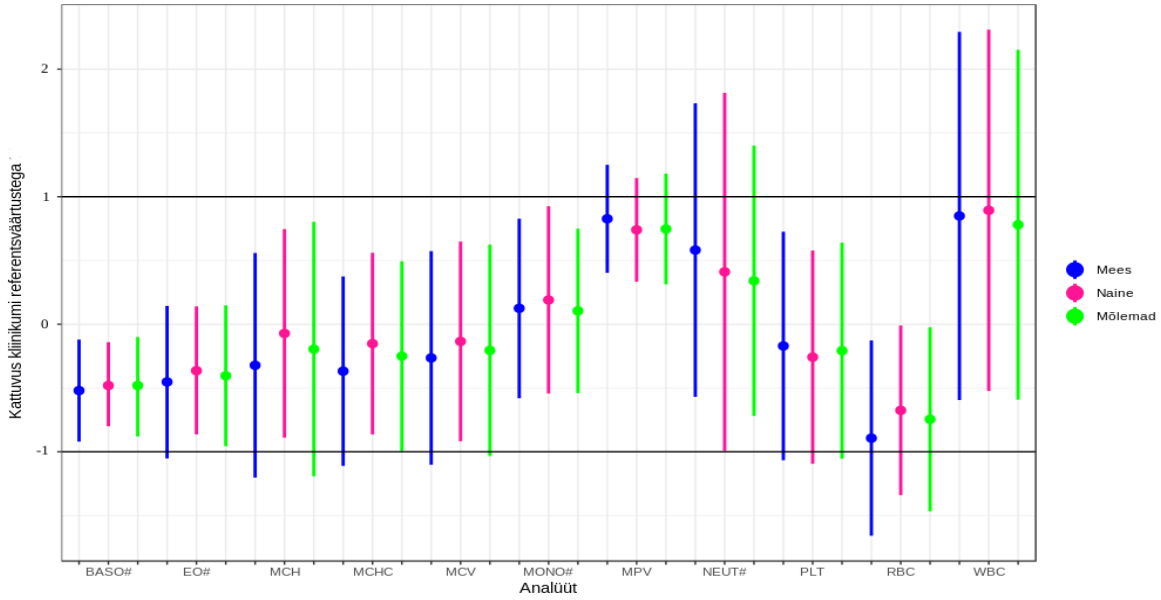
Jooniselt ilmneb, et üldjuhul on mudeli poolt leitud intervallid kitsamad ehk normaalne veri on rangemalt defineeritud. Põhjus võib peituda selles, et TÜK-is võib olla referentspiiri varieeruvus 2σ. Ühtegi põhimõttelist vastuolu TÜK-i ja mudeli poolt leitud vahemikes pole.

WBC, NEUT, MPV puhul on kliinikumi soovituslik ülempiir madalam kui mudeli poolt leitud ülemine piir ehk mudel peab normaalseks rohkem väärtusi kui TÜK. RBC, PLT ja MCH korral on mudeli alumine referentspiir madalam kliinikumi omast ehk mudel võib anormaalset verenäitajat lugeda normaalseks. Analüütide MCH,MCHC,MCV ja PLT puhul kattuvad mudeli leitud referentsvahemikud tegelikega väga suurel määral. Analüütide BASO# ja EO# mudelist saadud vahemikud on väga kitsad. Intervalli paiknemine nullist allpool näitab, et referentsvahemike alumised pooled kattuvad. Samas on juba algselt TÜK-i andmetel basofiilide ja eosinofiilide vahemikud väga väikesed, vastavalt 0-0.1 ja 0-0.5. Seega ka juba väiksel erinevusel mudeli ja kliinikumi vahel on suur mõju joonisele. Naiste ja meeste seas märkimisväärseid erinevusi ei ole.

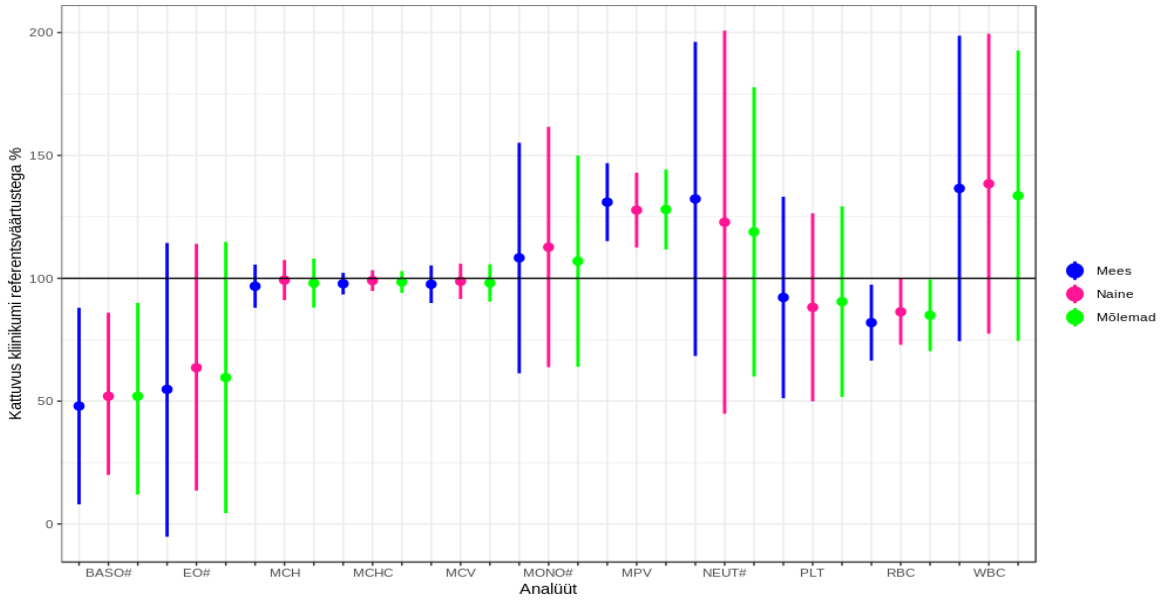
Joonis 7 näitab mudeli vahemike käitumist keskmiste suhtes protsentuaalselt. Jooniselt ilmneb, et tulemused on jällegi suurel määral kooskõlas TÜK-i vahemikega. Analüütide BASO#, EO#, MONO#, NEUT#, WBC# puhul on relatiivne varieeruvus suurem võrreldes TÜK-ga (umbes 100%). MCH, MCHC, MCV MPV ja RBC puhul on relatiivne varieeruvus väiksem (alla 10%) ehk lubatud bioloogiline variatsioon on nende analüütide puhul väiksem.

Lisaks vaatleme, kas referentsvahemikud erinevad vanusegrupiti. Selleks jagame patsiendid kümne aasta kaupa vanusegruppidesse – 0...9, 10...19, ..., 80...89, 90+ aastased. Normaalseks vereks defineerime mudeli poolt leitud kogu populatsiooni normaalsed näitajad. Leiame igale vanusegrupile kõikide verenäitajate $0.25, 0.5$ ja 0.75 kvantiilid. Kuna analüüte on kokku 11, siis tulemuste visualiseerimiseks kasutame soojuskaarti. Selleks on kõigi analüütide väärtused standardiseeritud mediaani järgi (valem (7)).

Jooniselt 8 ilmneb, et vanusegrupis 0...9 on BASO#,EO#, MONO#, PLT ja WBC väärtused kõrgemad võrreldes ülejäänud vanusegruppidega. Tulemust selgitab asjaolu, et TÜK-s on toodud alla 18 aastastele eraldi referentsvahemikud, mis on kõrgemad võrreldes täiskas-

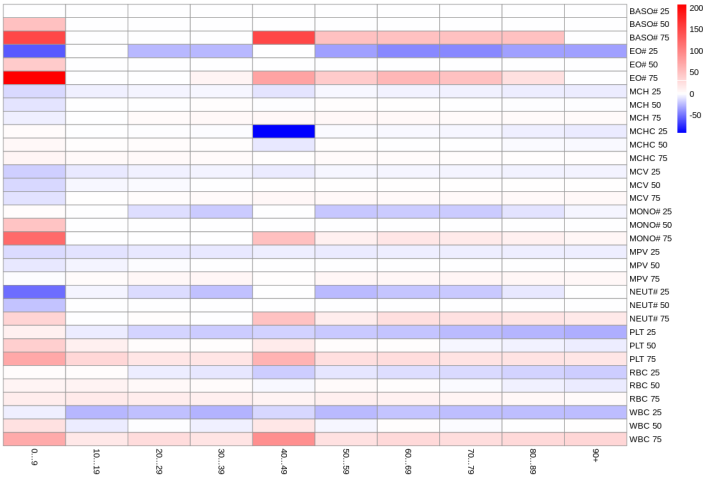


Joonis 6: Mudeli poolt leitud referentsvahemike arvuline kattuvus TÜK-i referentsvahemikega sugude lõikes



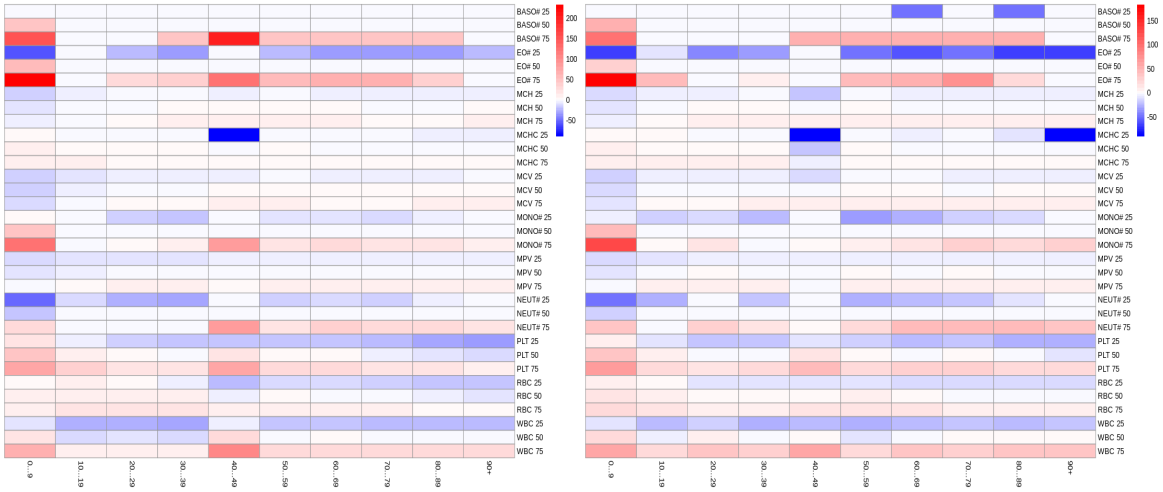
Joonis 7: Mudeli poolt leitud referentsvahemike protsentuaalne kattuvus TÜK-i referentsvahemikega sugude lõikes

vanutega. Silma paistab ka vanusegrupp 40...49, kus mitmete analüütide: BASO#, EO#, MONO#, NEUT#, PLT, WBC 0.75 kvantiilid on kõrgemad muudest vanusegruppidest. Selle üheks põhjustajaks võib meeste puhul olla metaboolse sündroomi teke ja naiste puhul menopaus. MCHC puhul aga esineb tunduvalt rohkem madalaid väärtusi alumises kvantiilis.



Joonis 8: Normaalse vere referentsvahemikud vanusegrupiti

Uurimaks andmeid lähemalt jaotame vanusegrupid soo järgi kaheks. Joonisel 9a on näha, et just meeste mediaanist kõrgemad analüüdi väärtused vanuses 40...49 mõjutasid joonise 8 sama vanusevahemikku. Naiste puhul domineerib soojuskaardil EO# ja BASO# juures sinine värv, viidates sellele, et naistel on tavaliselt meestest madalamad verenäitajad (joonis 9b). Võib järeldada, et laste puhul ei sõltu verepildi muutus soost, üle 40 aastaste puhul tulevad aga välja sugudevahelised erinevused.



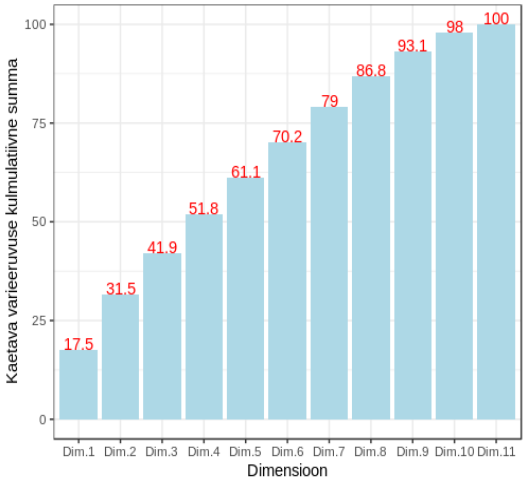
(a) Mehed (b) Naised

Joonis 9: Meeste ja naiste normaalse vere referentsvahemiked vanusegrupiti

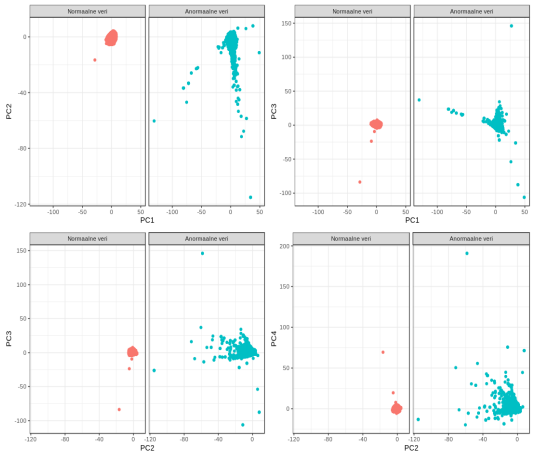
3.2 Peakomponentide analüüs

Klastrite eraldumise visualiseerimiseks rakendame peakomponentide analüüsi. Tuleb välja, et säilitamaks 95% andmete varieeruvusest tuleb alles jätta üheteistkümnest peakomponendist üheksa (joonis 10a). Seega ei saa PCA abil andmestiku dimensiooni väga suurel määral vähendada. Jooniselt 10b on näha, kuidas anormaalne veri on tunduvalt rohkem varieeruvate väärtustega võrreldes normaalse verega. Tulemus on loogiline, sest sõltuvalt haigusest võib anormaalset verd olla mitut erinevat tüüpi ja väärtused olla ekstreemsemad.

Tuleb märkida, et tegelikult ilmens juba eelneva peatüki korrelatsioonimaatriksis (joonis 5), et enamik tunnuseid on sõltumatud. PCA eelduseks on sõltuvad tunnused, seetõttu ei anna PCA meile väga informatiivset tulemust. Teiseks probleemiks on asjaolu, et anormaalse vere puhul on enamike analüütide väärtus paigas ja ainult üksikud väärtused erinevad tunduvalt normaalsest. Seega paigutuvad paljud anormaalse vere punktid joonisel normaalse klastri punktidega samas kohas.



(a) Peakomponentide varieeruvus



(b) Anormaalne ja normaalne veri kujutatud esimesel neljal peakomponendil

Joonis 10: Peakomponentide analüüs

3.3 Verepilti muutvad haigused

Järgneva alampeatüki eesmärgiks on normaalse ja anormaalse vere klasterdamise tulemuste valideerimine. Siinkohal rõhutame, et alampeatükis ei tuvastata uusi diagnostilisi meetodeid. Kontrolli sooritamiseks uurime, millised diagnoosid põhjustavad verepildi muutumist. Kuna uuritavad patsiendid valime välja juba varasemalt diagnoositud haiguse põhjal, siis on tegu retrospektiivse uuringuga. Võrdlemaks haiguste esinemist normaalses ja anormaalses veres kasutame juht-kontroll uuringut. Juhtgruppi kuuluvad kõik

anormaalse verega patsientide diagnoosid ja kontrollgruppi normaalse verega patsiendite diagnoosid. Juhtgruppi võtame kõik 332 426 anormaalse verega patsientidele määratud diagnoosid ja kontrollgruppi 2 585 319 normaalse verega patsientidele määratud diagnoosi.

Igale vereanalüüsile vastav diagnoos on tähistatud RHK-10 koodina. Diagnooside koodid on väga täpsed (näiteks A01.0 on kõhutüüfus), seega moodustame diagnoosigrupid RHK-10 alampeatükkide kaupa (näiteks A01.0 kuulub alampeatükki A00-A09 ehk soolenakkushaiguste alla), selleks et igasse gruppi jääks piisavalt vaatlusi. RHK-10 alampeatükke (edaspidi lihtsalt alampeatükk) on kokku 259. Tulemuseks on sagedustabel, kus on toodud iga diagnoosi esinemissagedus anormaalses ja normaalses veres.

Peamist huvi pakub meile diagnooside sageduse terves ja haiges veres, nende võrdlemiseks kasutame šansside suhet. Šansside suhet kasutame võrdlemaks kui mitu korda erineb vastava diagnoosi saamise šanss haige ja terve verepildiga inimestel. Tulemuste visualiseerimiseks kasutasime kokteiliklaasi joonist (*volcanoplot*) , mille tulemus on näha jooniselt 11. Joonise x-teljel on logaritmitud šansside suhted alusel 2. Punktide paiknemine x-telje nullpunktist vasakul pool osutab diagnoosi esinemisele sagedamini normaalses veres, paremale pool aga anormaalses veres. Logaritmitud šansside suhtel on tõmmatud vertikaalne katkendjooned kohta, kus šanss kuuluda ühte gruppi (juht/kontroll) on kaks korda suurem kui teise.

Võrdlemaks, kas oodatav ja tegelik sagedus erinevad, leiame iga diagnoosigrupi korral χ^2 testi abil p-väärtused, mis näitavad, kas sagedused juht- ja kontrollgrupis on statistiliselt erinevad. χ^2 testi kasutame ainult selleks, et eemaldada juhuslikkusega seletatavad efektid. Tabeli 259st diagnoosigrupist on 19 grupi esinemissagedus väiksem kui viis (C97,F99, H55-H59, P75-P78, Q10-Q18, Q30-Q34, R83-R89, R95-R99, V30-V39, V50-V59, V60-V69, V70-V79, V80-V89, V90-V94, V95-V97, V98-V99, W65-W74, W75-W84, Y35-Y36); nende gruppide puhul ei pruugi p-väärtus täpne olla.

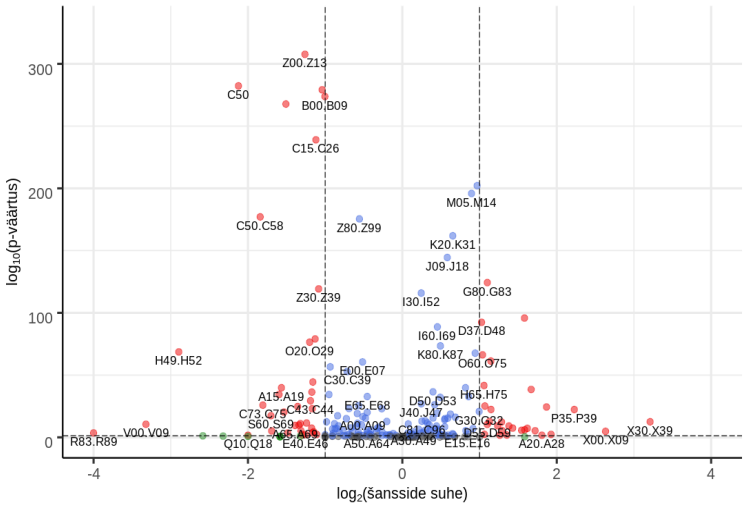
Joonise y-teljel on absoluutväärtus logaritmitud p-väärtustest alusel 10. Mida suurem on logaritmitud p-väärtus, seda statistiliselt olulisem on diagnoosi esinemissageduste erinevus juht- ja kontrollgrupis. Mitmese testimise korral korrigeerime lävendit $\alpha = 0.05$, saades uueks lävendiks

$$\alpha' = \frac{0.05}{\text{paariviisi testide arv}} \ .$$

Horisontaalne katkendjoon on joonistatud vastavalt $|(\log_{10}(\alpha')|$ juurde.

Jooniselt pakuvad huvi eeskätt punast värvi punktid. Nende diagnooside puhul on nii

šansside suhe kui ka logaritmitud p-väärtus suur. Järelikult on nende diagnooside esinemise šanss anormaalses veres üle 2 korra suurem võrreldes normaalse verega ja diagnooside esinemissageduste erinevus on statistiliselt oluline.



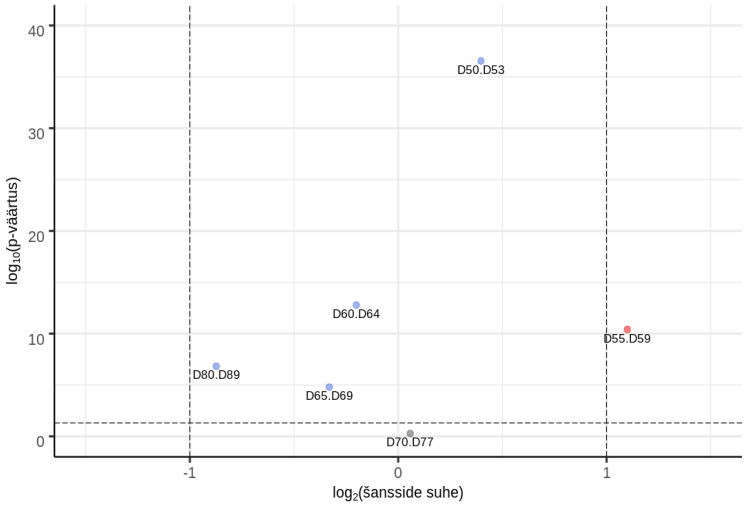
Joonis 11: Kõik RHK-10 klassifikaatorite alampeatükid

Oluline on märkida, et tegelikkuses ei saa me jooniste põhjal öelda, et näiteks angiin põhjustab verepildi muutumist. Mingi kolmas faktor võib esile kutsuda nii verepildi muutumist kui ka angiini. Et väita põhjuslikkust, peaksime moodustama juhtgrupi kõigi angiinihaigetega ja kontrollgrupi juhtgrupile sarnastest inimestest, kellel pole angiini. Edaspidisel jooniste tõlgendamise lihtsustamisel eeldame, et eelnev informatsioon on teada ega rõhuta eraldi korrelatsiooni erinevust põhjuslikkusest.

Vaatleme eraldi RHK-10 klassifikaatoreid (täpsemalt toodud tabelis 3), mille puhul võib oletada, et diagnoos peaks muutma verepilti. Huvi pakuvad järgnevad RHK-10 peatükid: peatükk III „Vere- ja vereloomeelundite haigused ning teatavad immuunmehhanismidega seotud haigusseisundid", peatükk X „Hingamiselundite haigused", peatükk I „Teatavad nakkus- ja parasiithaigused" ja peatükk XXI „Terviseseisundit mõjustavad tegurid ja kontaktid terviseteenistusega". Peatükkide alla käivate diagnooside lähemalt uurimiseks teeme iga peatüki kohta kokteiliklaasi joonised.

Joonisel 12 on välja toodud „Vere- ja vereloomeelundite haigused ning teatavad immuunmehhanismidega seotud haigusseisundite" diagnoosid. Kõige eristuvamateks diagnoosigruppideks on „Hemolüütilised aneemiad" (D55-D59) ja „Toitumisaneemiad" (D50-53), mille diagnoose on anormaalses veres rohkem võrreldes normaalse verega. Perearst Evelin Raie sõnul peab tulemus paika, sest aneemiat diagnoositaksegi tavaliselt punaverenäi-

tude (MCV, EO#, MCHC, MCH, MPV) põhjal [13]. „Teatavad immuunmehhanismi haaravad haigusseisundid“ (D80-D89) tema sõnul vereanalüüside tulemustes tingimata ei ilmne, mis on kooskõlas meie joonisega (paikneb vasakul pool ehk esineb rohkem normaalses veres).

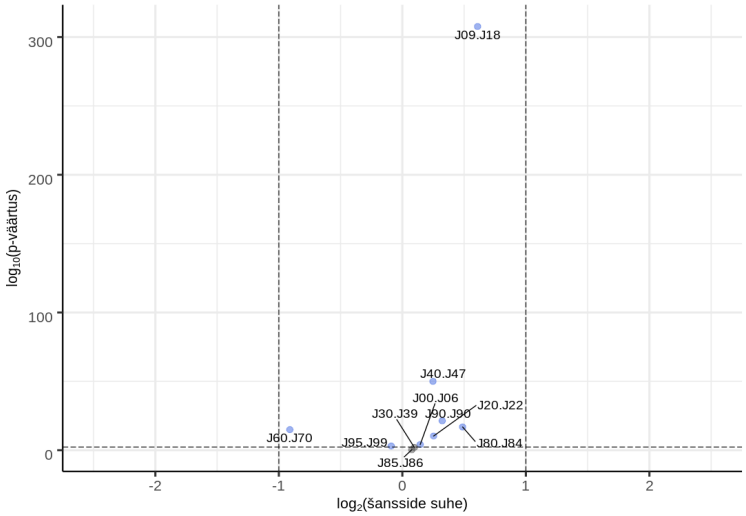


Joonis 12: Vere- ja vereloomeelundite haigused ning teatavad immuunmehhanismidega seotud haigusseisundid

Joonisel 13 on välja toodud peatüki „Hingamiselundite haiguste“ (J00-J99) alampeatükid. „Gripp ja pneumoonia e kopsupõletik“ (J09-J18) sagedus normaalses ja anormaalses veres eristub märkimisväärselt teistest alampeatükkidest. Raie sõnul on kopsupõletike puhul nihked verenäitajates sagedased. Näiteks tõusevad bakteriaalse kopsupõletiku (J15) puhul eeskätt WBC ja NEUT# arvud. „Välistegurite põhjustatud kopsuhaigused“ (J60-J70) juures märkis Raie, et need pole väga tihti esinevad diagnoosid vaid pigem kroonilise iseloomuga, seega ei pruugigi nende puhul olla verepildis suuri muutusi. [13]

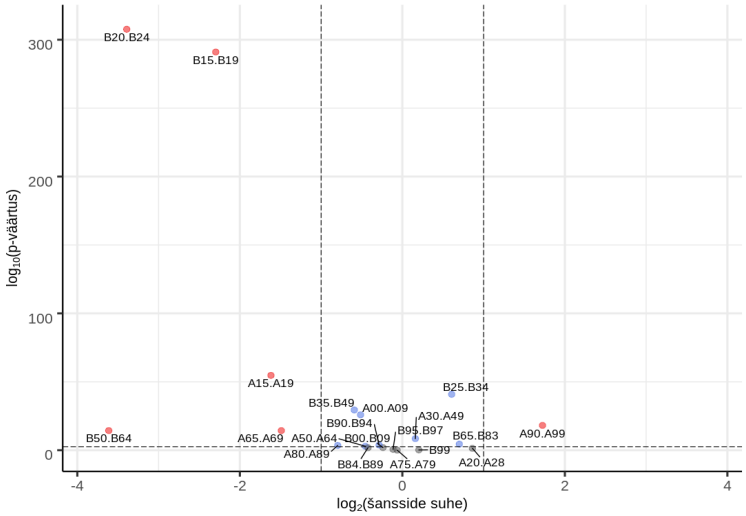
„Teatavate nakkus- ja parasiithaiguste“ (A00-B99) joonisel 14 paistavad välja, et normaalses veres esinevad sagedasemalt diagnoosid „Viirushepatiit“(B15-B19), „Inimese immuunpuudulikkuse viirustõbi e HIV-tõbi“ (B20-B24) „Algloomhaigused“ (B50-B64) ja anormaalses veres „Lülijalgse levitatavad viiruspalavikud ja hemorraagilised viiruspalavikud“ (A90-A99).

HIV diagnoosi puhul on üllatav tulemus, et diagnoosil on $\log_2(x) \approx 3.7$ ehk 13 korda suurem šanss esineda normaalses veres (joonisel vasakul paiknemine näitab, et HIV ei mõjuta palju verepilt). Järelikult elavad HIV diagnoosiga patsiendid enamasti normaalset elu seni kuni haigus on kontrolli all. Raie toob välja, et HIV ei mõjutagi tingimata verenäi-



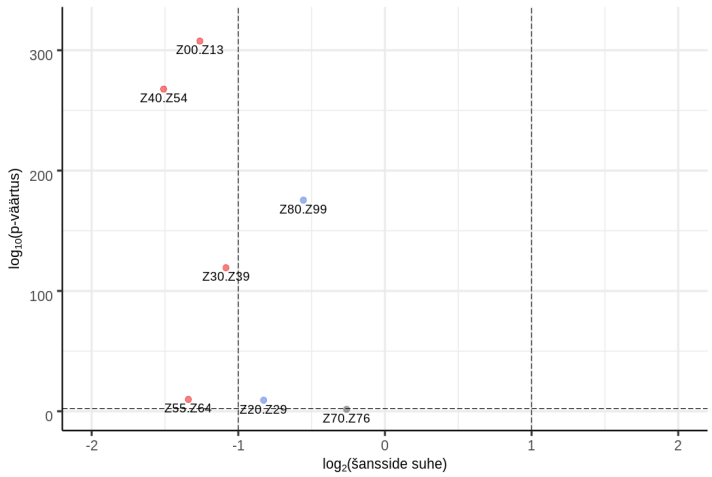
Joonis 13: Hingamiselundite haigused

tajaid ja veri võib olla normaalne, küll aga ravi toimimise uurimiseks võetakse tihti vereproove kontrollimaks, kas leukotsüüdid (WBC) püsivad normi piires. Jooniselt on näha, et viirushepatiidiga diagnoositud patisentide verenäitajad püsivad normaalsed. Raie sõnul viirushepatiidid verenäitajates tavaliselt ei kajastugi. Algloomhaiguste tulemus on mõneti üllatav, sest sinna alla kuulub malaaria (B50-B54), mis peaks WBC ja RBC näitajaid muutma, jooniselt on aga näha, et selle haiguste grupi diagnoositutel on verepilt jällegi 13 korda suurema šanssiga korras. Raie hinnangu kohaselt peaksid viiruspalavikkude puhul verenäitajad muutuma, mis on kooskõlas joonisega. Näiteks dengue palaviku (A90 ja A91) puhul on RBC ja WBC näitajad normaalsest madalamad [13].



Joonis 14: Teatavad nakkus- ja parasiithaigused

Joonisel 15 on visualiseeritud diagnoosigruppi „Terviseseisundit mõjustavad tegurid ja kontaktid terviseteenistusega“.Peatükki alla kuuluvad kõik tervisekontrolli eesmärgil tehtud uuringud. Kuna enamik tervisekontrollis käivaid inimesi on tegelikult terved (lihtsalt näiteks töö või kool kohustab tervisekontrollis käimist), siis vastab joonis ka oodatavale ehk normaalse verepildiga patsientidel esineb tervisekontrolli diagnoose tunduvalt rohkem kui haige verepildiga patsientidel.



Joonis 15: Terviseseisundit mõjustavad tegurid ja kontaktid terviseteenistusega

4 Anormaalse vere klasterdamine

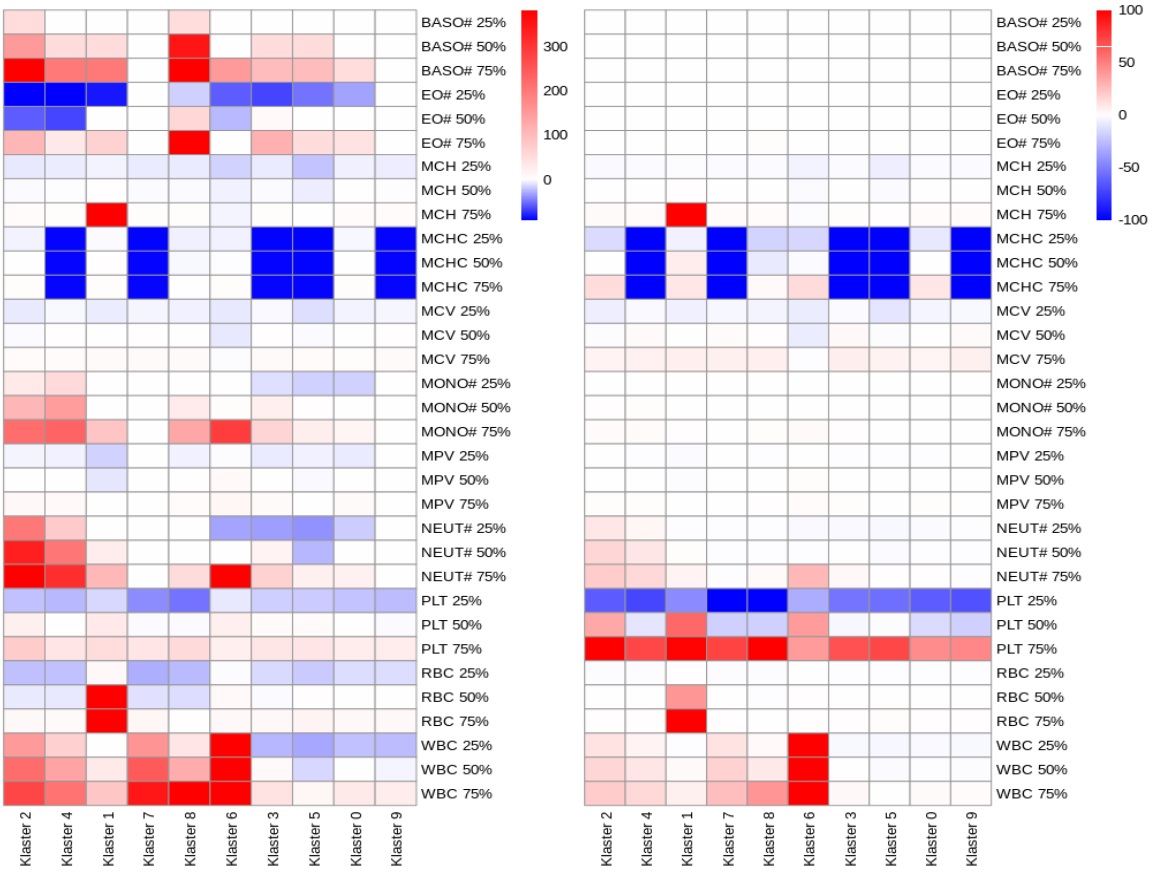
Haige patsiendi veri pole alati ühesugune. Mõni haigus muudab ühe analüüdi väärtust, mõni haigus teise, näiteks aneemiale viitab madal hemoglobiini tase, külmetuse puhul on aga valgete vereliblede arv tavalisest madalam. Kasutades Gaussi segujaotust klasterdame kõik anormaalsed vere mõõtmised. Segumudel leiab ka optimaalseima klastrite arvu, milleks on 9 klastrit. Tabelis 2 on välja toodud, mitu vaatlust sattus igasse klastrisse.

Tabel 2: Anormaalse vere klastrite suurused

Klastri number	1	2	3	4	5	6	7	8	9
Vaatluste arv	83	30 828	35 649	2 561	4 611	37	3 246	26 276	228 421

Kõigile klastritele on arvutatud ka referentsvahemikud samamoodi nagu peatükis 3.1. Referentsvahemike paremaks hoomamiseks on anormaalse vere referentsvahemikud kujutatud soojuskaardi joonisel 16a, kus on näha anormaalse vere klastrite analüütide 0.25, 0.5 ja 0.75 kvantiilid. Joonisel tähistab klaster 0 normaalset verd ja klastrid 1-9 erinevat tüüpi anormaalset verd. Kuna analüütide väärtused on väga erinevas suuruses (näiteks on BASO# referentsvahemik 0-0.1, MCHC aga 317-357), siis on kõigi analüütide väärtused standardiseeritud normaalse vere mediaani järgi (valem (7)). Joonisel on näha, millised verenäitajad igas klastris normaalsest erinevad. Ruudu valge värvus näitab, et analüüdi väärtus vastab normaalsele verele. Klastrites $2,6,7,8$ on WBC näitajad tavalisest kõrgemad. Klastrites $2,8$ on BASO# ja klastrites $2,4$ on NEUT# näitajad kõrged. Normaalsest madalamad väärtused paistavad silma MCHC puhul klastrites $3,4,5,7,9$ ja EO# puhul klastrites $2,4$. Kvantiilide vaheline erinevus (0.25, 0.5, 0.75) näitab klastrisisest varieeruvust, näiteks klastris 6 varieerub NEUT# arv suurel määral.

Kasutasime samadel andmetel ka klassikalist standardiseerimist (joonis 16b, standardiseerimise valem (6)). Silma paistavad peamiselt analüüdid PLT ja MCHC, kuna nende normaalne bioloogiline varieeruvus on suurem võrreldes teiste analüütidega. Näiteks on MONO# normaalne väärtuste vahemik 2-10, aga MCHC puhul 317 - 357 ja PLT puhul 145 - 390.



(a) Mediaani järgi standardiseerimine

(b) Klassikaline standardiseerimine

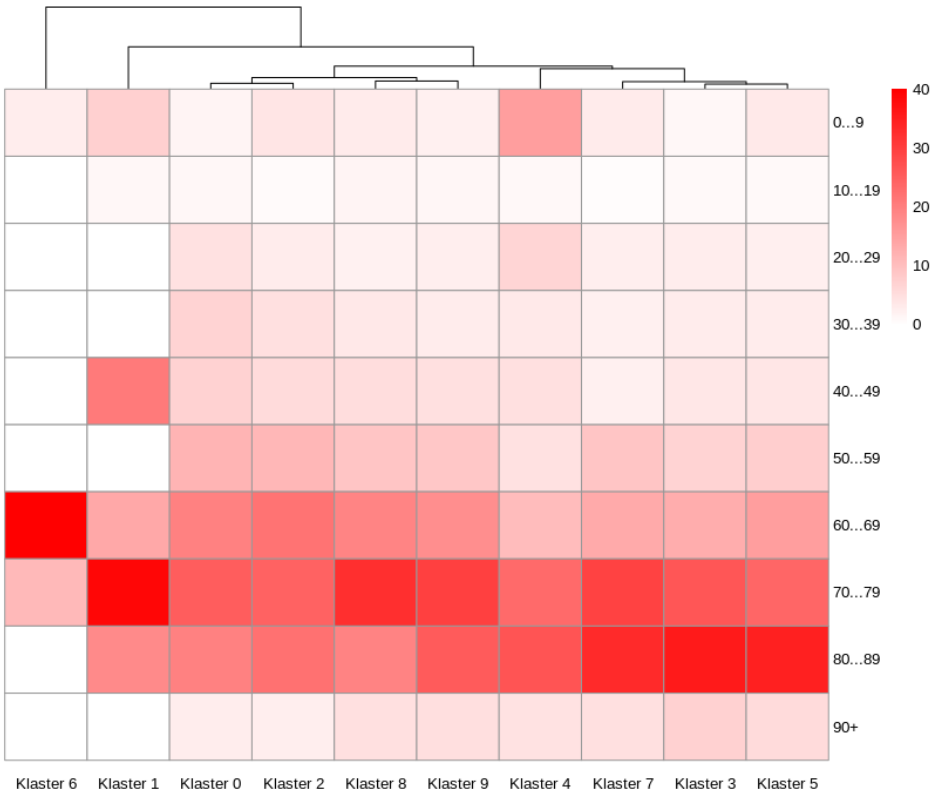
Joonis 16: Anormaalse vere analüütide referentsvahemike kvantiilid

4.1 Anormaalse verepildi vanuseline jaotuvus

Järgnevalt võrdleme anormaalse vere klastrite vanuselist jaotuvust. Klastrite suurused on väga erinevad, seega leiame iga vanusegrupi protsentuaalse osakaalu klastris. Saadud tulemused on toodud joonisel 17. Valged ruudud näitavad, et vastava vanusegrupi patsiente klastris ei esine. Üldpildis paistab klastrite vanuseline jaotuvus sarnane. Kõikides klastrites moodustavad suurima osa üle 60 aastased patsiendid (tugevamad punased toonid). Tulemus on loogiline, sest vanemad inimesed käivad tavaliselt rohkem arsti juures. Klaster 4 paistab välja teistest klastritest suurema laste osakaalu poolest.

4.2 Trajektoorid

Uurimaks keerulisi bioloogilisi protsesse on mõõtmiste diskretiseerimine väga levinud viis, kuna ajaline evolutsioon 11-mõõtmelises ruumis (uurime 11 vereanalüüti) pole visualiseeritav ega jälgitav. Vaatleme, kuidas patsiendi verepilt liigub ajas erinevate klastrite vahel kuni patsiendi vere normaliseerumiseni. Selleks koostame igale patsiendile vere-



Joonis 17: Vanusegruppide osakaalud anormaalse vere klastrites

näitajate muutumise trajektoori.

Trajektooride moodustamisel jätame alles patsiendid, kellelt on vähemalt kaks korda võetud vereproovi. Trajektoor alguspunktiks on aeg, kus patsiendil esimest korda diagnoositakse haigus (alguspunktiks ükskõik milline klaster), iga järgneva mõõtmise aeg on päevade arv esmasest diagnoosist. Trajektoor lõpeb kui patsiendi veri jõuab tagasi normaalsesse olekusse ehk klastrisse 0. Esineb olukordi, kus trajektoor ei lõppe normaalse verega, kuigi näiteks on äärmiselt ebatõenäoline, et angiin kestab mitu aastat. Sellisel juhul määrame käsitsi maksimaalse ajalise trajektoori pikkuse, pärast mida loeme trajektoori lõppenuks ja järgnevat mõõtmist uue trajektoori alguseks.

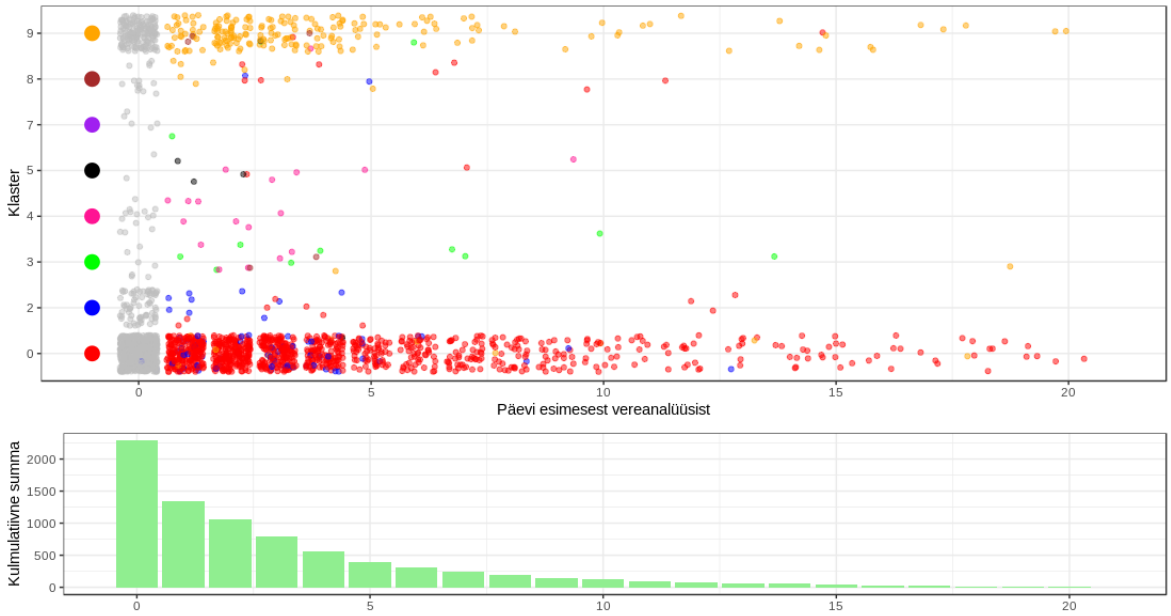
4.2.1 Haiguspõhine trajektoor

Võtame vaatluse alla angiini (RHK-10 koodid J03 ja J35) ja aneemia (RHK-10 koodid D50-D53, D55-D59, D60-D64) diagnoosiga patsiendid, sest need diagnoosid peaksid muutma verenäitajaid. Uurime nende verepildi muutumise trajektoore haiguse jooksul.

Angiini haigestunud patsiendile määratakse tavapäraselt raviks antibiootikumide kuur kestusega üks kuni kaks nädalat [14]. Seega määrame maksimaalseks trajektoori pikku-

seks nädalase varuga 21 päeva. See tähendab, et kui patsiendil haigus on kestnud kauem kui 21 päeva, loeme selle juba uueks angiiniks.

Joonisel 18 on toodud angiini trajektoorid. Suur täpp näitab, mis värviga vastav klaster on tähistatud. Väikeste täppide värvused näitavad eelnevat klastrit ehk mis klastrist on jõutud käesolevasse klastrisse. Kui joonise väikesed punktid on sama värvi, mis suur punkt, siis näitab see seda, et verepilt on püsinud samas klastris. Näiteks klastris 9 (klastri värv oranž) trajektooris on samuti väga palju oranže väikeseid täppe, mille põhjal saab järeldada, et enamjaolt püsib veri samas klastris. Huvi pakub klaster 0 ehk terve vere klaster: seal on näha peale punase ka siniseid, pruune ja oranže täppe. See tähendab, et kui haiguse alguses on veri 2, 8 või 9 tüüpi anormaalne, siis on patsiendile suur võimalus jõuda normaalsesse olekusse. Histogrammil on näha trajektoori pikkuste kumulatiivne summa. Esimene tulp tähistab kõigi mõõtmiste arvu, kus trajektoori pikkus oli vähemalt 0 päeva, teine tulp, kus trajektoori pikkus oli vähemalt 1 päev jne. Histogramm selgitab, miks trajektooride joonisel punkte aina hõredamaks jääb. Võib järeldada, et angiin kestab harva üle 10 päeva.



Joonis 18: Angiini trajektoor

Aneemia puhul pole maksimaalse ajalise trajektoori pikkuse määramine nii lihtne, sest tervenemine ei sõltu ainult ravimitest, vaid ka inimeese toitumisest, mille muutmine on pikaajalisem protsess. Seega on võetud trajektoori maksimaalseks pikkuseks 365 päeva. Jooniselt 19 ilmneb, et peamiselt püsib vereanalüüsi tulemus samas klastris. Normaalses veres (klaster 0) on rohkem erinevat värvi täppe (põhiliselt sinised, pruunid ja oranžid); järelikult patsiendid, kelle verenäitajad on eelnevalt olnud klastrites 2, 8, 9, saavad kõige tõenäolisemalt terveks. Paljudel juhtudel pole patsiendile viimast „veendume, et on ter-

ve“ mõõtmist tehtud ja see võib olla põhjuseks, miks ülejäänud anormaalse vere klastrist normaalsesse ei liigu. Histogrammilt on näha, et kumulatiivne summa trajektooride pikkustest kukub kiirelt.

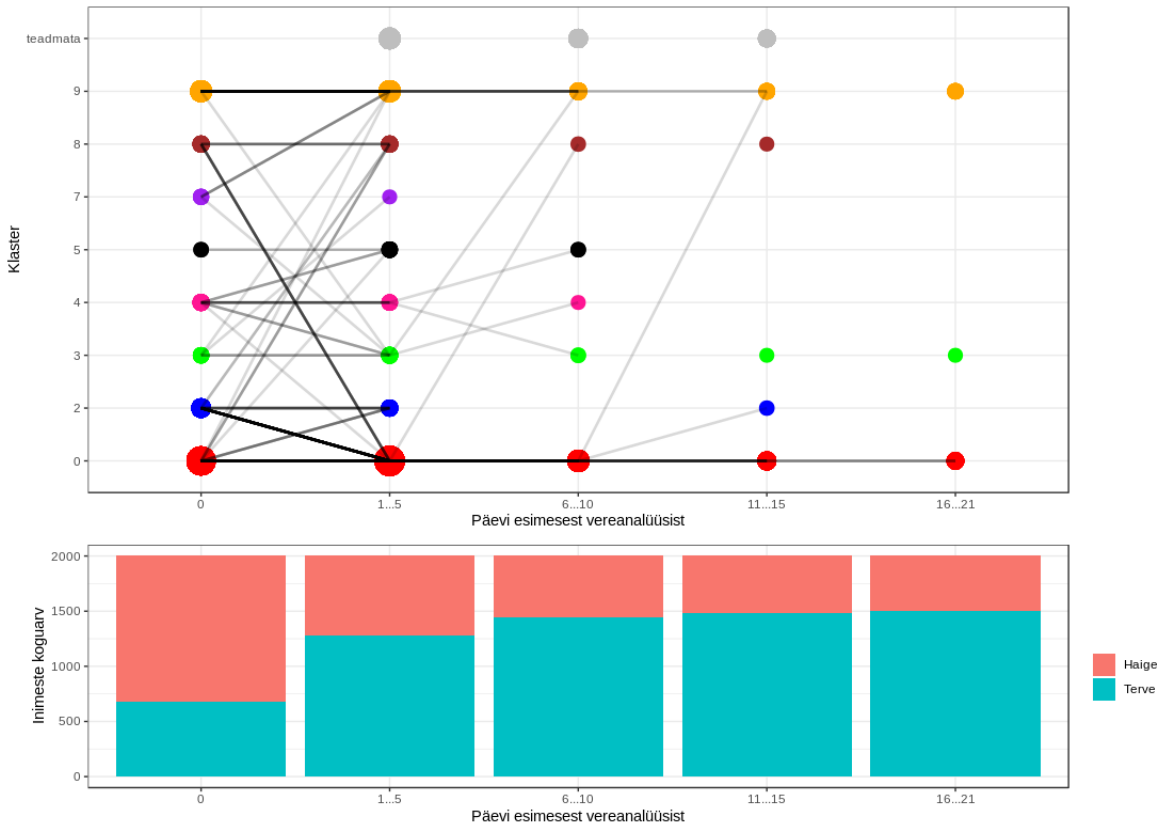
Joonis 19: Aneemia trajektoor

4.2.2 Haiguspõhine trajektoor diskreetse ajaga

Eelnvalt vaatlesime angiini trajektoori 21 päeva jooksul ja aneemia trajektoori 365 päeva jooksul pärast esimest vereanalüüsi. Nägemaks täpsemalt klastrite vahel liikumise trajektoore teeme ajast järjestustunnuse.

Jaotame angiini trajektoori viie päeva kaupa gruppidesse vastavalt 1...5, 6...10, 11...15 ja 16...21 päeva pärast esimest vereanalüüsi. Kui ajavahemiku jääb mitu väärtust võetakse neist esimene. Joonisel 20 on toodud välja patisendi liikumine ajagruppide vahel. Joone tugevus vastab trajektoori sagedusele ja täpi suurus näitab, kui tihti on patsientide veri vastavas ajavahemikus vastavas klastris. Osale patsientidest ei tehta veremõõtmisi iga 5 päeva tagant. Kui patsiendil pole vastavas ajavahemikus vereanalüüsi tehtud, siis kuulub ta klassi „teadmata“ alla (hallid punktid joonisel). Punktid, mis pole ühendatud ühegi joonega, tulevadki „teadmata“ klassist, kuid joonise selguse säilitamise eesmärgil on jäetud need jooned joonistamata.

Jooniselt ilmneb, et esimese viie haiguspäeva jooksul on vereanalüüdid väga muutlikud. Kõigis klastrites (v.a 7) võib jääda verepilt samaks (horisontaalsed jooned, päevast 0 päevadeni 1...5) ehk arvatavasti pole ravimid veel mõjuma hakanud või verepilt muutub muud



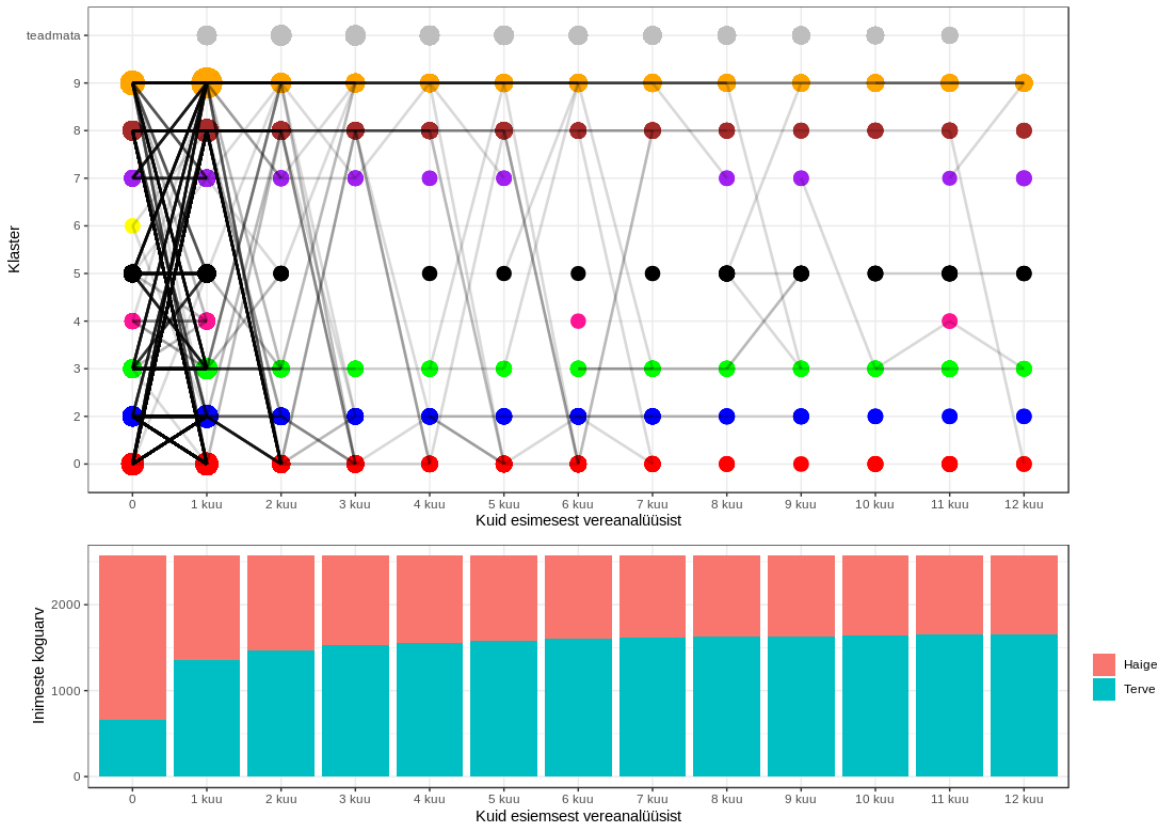
Joonis 20: Angiiniga patsientide verepildi muutumise trajektoorid

tüüpi anormaalseks ehk haigus areneb. Patsientidel, kelle veri esimesel mõõtmised oli klastris 2 või 8, on suur võimalus vähem kui nädalaga terveks saada. Leidub ka neid (klastrid 3, 4, 7), kelle verepilt esimesel viiel päeval küll muutub, kuid mitte terveks, vaid muud tüüpi anormaalseks. Hulk trajektoor saab alguse normaalsest verest ja seejärel liiguvad anormaalsesse, tegu võib olla patsientidega, kellel haigus aja möödudes hoopis ägeneb.

Järgneva 5 päeva jooksul (6...10) verepilt enam tüüpiliselt samasse klastrisse ei jää ja muutub teist tüüpi anormaalseks (klastrid 3 ja 4). Kuna igast ajagrupist on võetud esimene mõõtmine, siis on võimalik, et osa patsiente liigub haigest seisundist tervesse ajagrupi sees, kuid see joonisel ei kajastu.

Joonisel olevalt histogrammilt ilmneb, et enamiku inimeste verenäitajad saavad korda juba esimese viie päevaga. Histogrammil on võetud arvesse ka tervenemised, mis toimuvad ajagrupi sees. Tegelikkuses võib trajektoori alguspunkt olla varasem, sest diagnoos võib olla pandud alles mõni päev pärast haiguse tegelikku algust. Teatud osa inimesi ei tervene ka 3 nädala möödudes. Põhjuseks võib olla, et nende patsientide normaalne veri erineb-ki tunduvalt tavalise inimese verest või patsiendi tervenedes enam analüüsi ei tehta, mis kinnitaks, et inimese veri on jõudnud normaalsesse olekusse.

Aneemia puhul võtame patsiendi trajektoori pikkuseks ühe aasta, mille jaotame kuude



Joonis 21: Aneemiaga patsientide verepildi muutumise trajektoorid

kaupa gruppidesse. Jooniselt 21 on näha, et kõige sagedamased verepildi muutused toimuvad just esimesel kahel kuul. Verenäitajate muutusi võib toimuda ka ühe kuu sees, kuid need hetkel huvi ei paku ja seega on alles jäetud igast kuust selle kuu esimene mõõtmine.Tugevad horisontaalsed jooned klastrites 8 ja 9 näitavad, et nendes klastrites püsitakse pikalt ja stabiilselt. Enne poole aasta möödumist tervenevad patisendid, kelle veri on varasemalt olnud klastris $2, 8, 9$. Pärast 7 kuud ei vii normaalse vere klastrisse enam ühtegi joont, mis tähendab, et normaalsesse verre jõutakse läbi teadmata klastri või kui veri on varasemalt jõudnud normaalsesse olekusse, loeme trajektoori lõppenuks ja rohkem seda ei kajasta.

Samas on tervenenute osakaalu näitaval histogrammil võetud arvesse ka kõik kuu sees toimunud tervenemised. Aneemia tervenenute histogrammi vaadates paistab jällegi välja, et kõige rohkem tervenetakse just esimese kahe kuuga. Kolmandiku patsientide veri jääb ka pärast aastat aega anormaalsesse olekusse. Tulemust võib seletada sellega, et kui patsient tunneb ennast juba tervena, siis ei lähe ta alati uuesti vereanalüüsi tegema, et veenduda, kas ta tõesti on terve. Seega pole meil selliste patsientide puhul signaali tervenemisest, kuigi tegelikkuses on inimene saanud terveks.

Andmete esitamise probleemiks on see, et igast ajagrupist võetakse vaid esimene väärtus

ja seega läheb aja diskretiseerimisel osa informatsioonist kaduma. Samuti võib arst teha aneemia puhul analüüse kahe kuuse intervalliga, kuid see praegusel joonisel ei kajastu. Tulevastes analüüsides võiks puuduvate ajagrupi mõõtmiste korral eeldada, et veri lihtsalt püsib samas klastris.

4.2.3 Klastrisse kuulumise tõenäosus angiini korral

Kõigi klastrite vahelised piirid pole alati väga selged, sest mõned klastrid on üksteisega sarnased. Seega tahame leida, kui suure tõenäosusega kuulub patsiendile vastav vaatlus $\boldsymbol{x}_i$ klastritesse $0, 1, ..., 9$. Vaatluse alla on võetud angiini diagnoosiga patsiendid.

Klastrisse kuulumise tõenäosused saame leida kasutades järeltõenäosuseid γ_{ik}. Kuna Gaussi segumudelid on koostatud kahe sammuga (kõigepealt klasterdame normaalseks ja anormaalseks ja seejärel anormaalset verd veel omakorda), siis me ei saa kasutada otse paketi *mclust* leitud järeltõenäosused γ_{ik} vaid peame sisuliselt kordama EM-algoritmi E-sammu kombineerides kahe mudeli parameetrid.

Meil on teada kõigi klastrite keskväärtused $\boldsymbol{\mu}_0, ..., \boldsymbol{\mu}_9$ ja kovariatsioonimaatriksid $\boldsymbol{\Sigma}_0, ..., \boldsymbol{\Sigma}_9$. Teades kõikide klastrite suurusi saame leida ka segu kaalud

$$\pi_k = \frac{n_k}{n_0 + ... + n_9}, k = 0, ..., 9 \ .$$

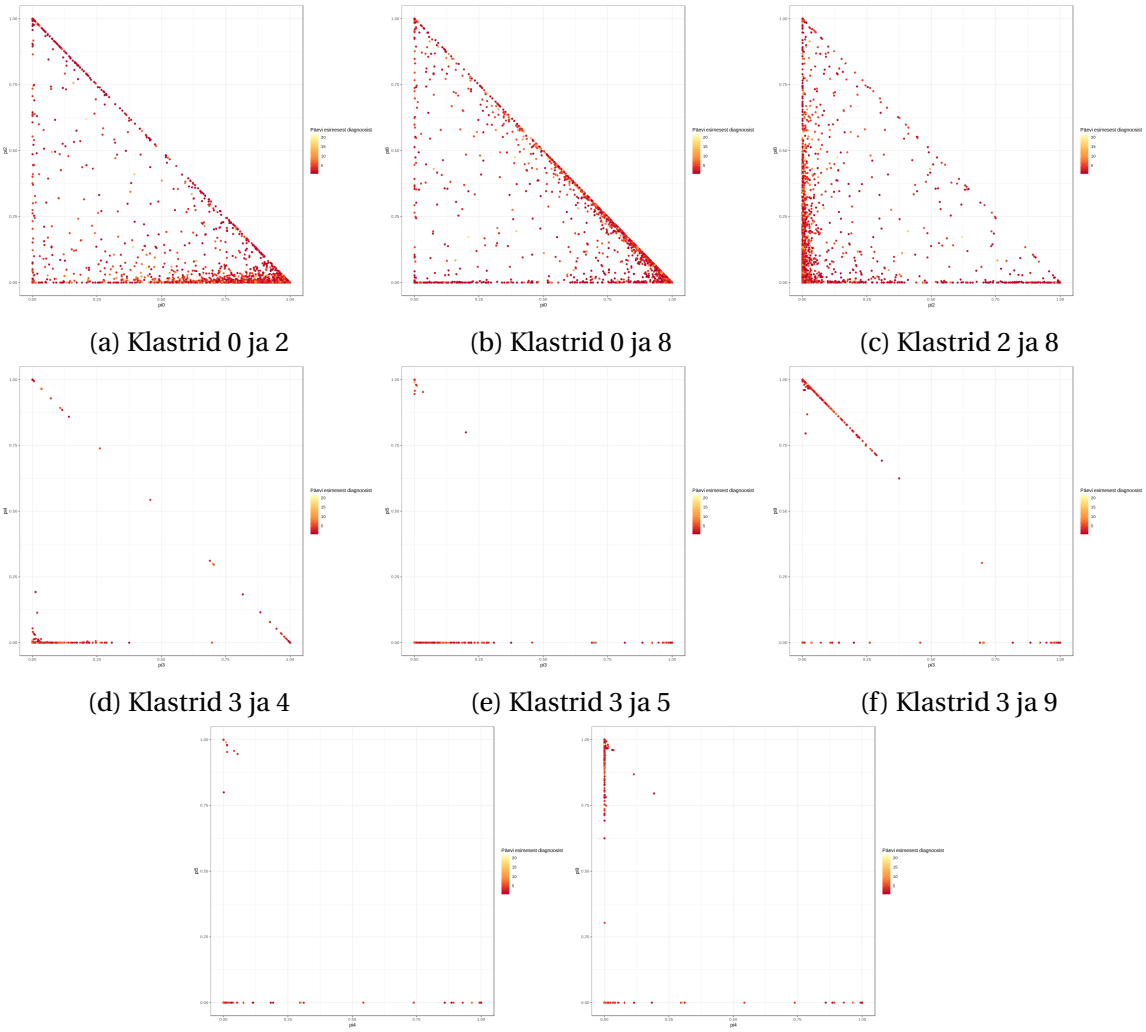
Seega on meil olemas kõikide parameetrite väärtused, et leida iga klassi $k = 0, ..., 9$ iga vaatluse i jaoks järeltõenäosus

$$\gamma_{ik} = \frac{\pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\pi_0 f_0(\boldsymbol{x}_i|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + ... + \pi_9 f_9(\boldsymbol{x}_i|\boldsymbol{\mu}_9, \boldsymbol{\Sigma}_9)} \ .$$

Praktikas on probleemiks, et lugeja või nimetaja väärtus on ülimalt väike ja seega pole arvutamine numbriliselt stabiilne. Selle lahendamiseks võib lugeja läbi jagada vaatluse i kõige tõenäolisema klassi k kuulumise normeerimata järeltõenäosusega $\pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Tehniliselt on antud arvutust kasulik teha logaritmilisel skaalal, sest siis on võimalike arvuti poolt leitavate väärtuste vahemik oluliselt suurem.

Visualiseerides tulemusi klastrite paaride kaupa (angiini diagnoosil on esindatud klastrid $0, 2, 3, 4, 5, 7, 8, 9$) tuleb jooniseid kokku 28. Joonisel 22 on välja toodud ainult mittetriviaalsed joonised. Joonise x-teljel on näha, kui suure tõenäosusega kuulub punkt x-teljel olevasse klastrisse ja y-teljel, kui suure tõenäosusega kuulub punkt y-teljel olevasse klastrisse. Kui punktid asuvad joonise keskel või diagonaalil, siis see näitab, et klastrid min-

gil määral kattuvad üksteisega. Kui punkte diagonaalil ei paikne, siis viitab see sellele, et klastrite vahel seoseid pole. Paiknemine peadiagonaalil näitab, et vaatlus kuulub kindlasti kas klastrisse i või klastrisse j, ja mitte kuhugi mujale. Kõige tugevamad seosed paistavad silma klastrite $0, 2, 8$ vahel.



(a) Klastrid 0 ja 2 (b) Klastrid 0 ja 8 (c) Klastrid 2 ja 8

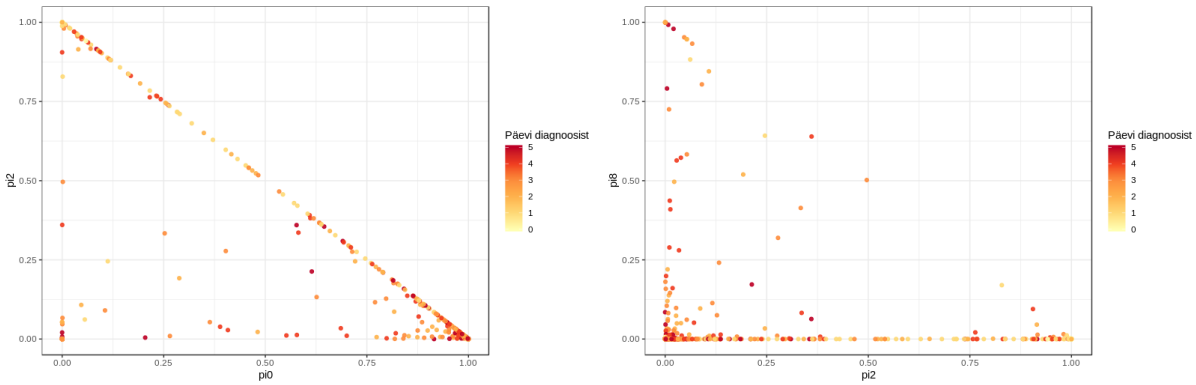
(d) Klastrid 3 ja 4 (e) Klastrid 3 ja 5 (f) Klastrid 3 ja 9

(g) Klastrid 4 ja 5 (h) Klastrid 4 ja 9

Joonis 22: Klastrisse kuulumise tõenäosus

Eelnevalt kirjeldatud jooniselt 22 ilmneb, et veri liigub peamiselt klastrite $0, 2$ ja 8 vahel. Seda kinnitab ka eelneva alampeatüki angiini joonis 20, kus nende klastrite vahel olid tugevad tumedad jooned. Samuti oli näha, et vere peamine liikumine klastrite vahel toimub esimesel viiel päeval. Vaatleme lähemalt, mis toimub esimesel viiel päeval trajektooridega, mis algavad klastrist 2 ja liiguvad klastritesse 0 või 8. Vastavad klastrisse kuulumise tõenäosused on kujutatud joonisel 23. Klastrist 2 liikumisel klastrisse 0 (joonis 23a) ilmneb, et esimestel päevadel pärast diagnoosi kuulub patsiendi veri kas normaalsesse verre või klastrisse 2 (ja mitte kuhugi mujale). Joonise diagonaalil tumedamad punktid on koondunud paremasse alumisse nurka. See näitab, et iga järgneva päevaga väheneb tõenäo-

sus kuuluda klastrisse 2 ja kasvab tõenäosus terveneda. Trajektoorid klastrist 2 klastrisse 8 enam nii tavapärased pole (joonis 23b), sest peadiagonaalil on punkte pigem vähem. See viitab sellele, et alustades klastrist 2 on tõenäosus kuuluda klastrisse 8 väiksem. Eelneva peatüki aja diskretiseerimisel ei ilmnenud, kui tõenäoline on kuuluda klastrisse 8. Jooniselt tulebki välja, et tegelikkuses määratakse klastrisse 8 ka punkte, mille tõenäosus kuuluda klastrisse 8 pole 1.



(a) Klastrid 0 ja 2

(b) Klastrid 2 ja 8

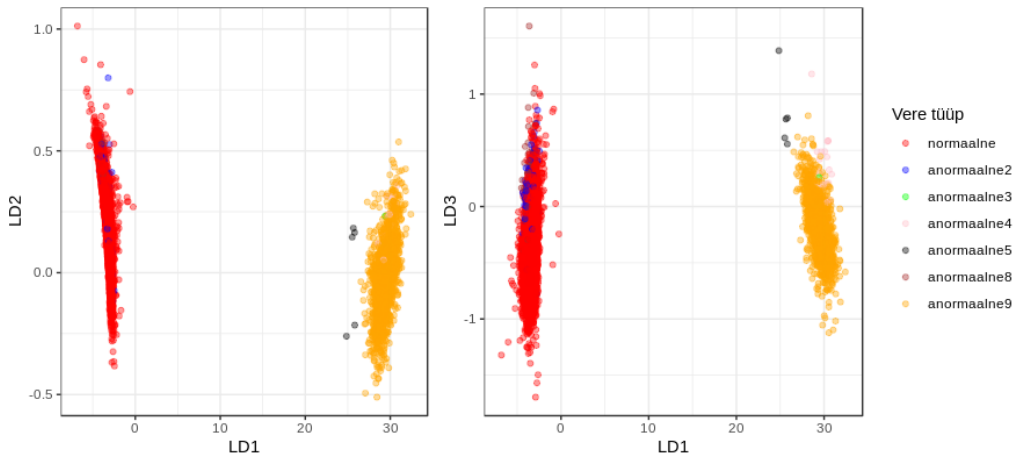
Joonis 23: Klastrist 2 algavate trajektooride klastrisse kuulumise tõenäosused

4.3 Lineaarne diskriminantanalüüs

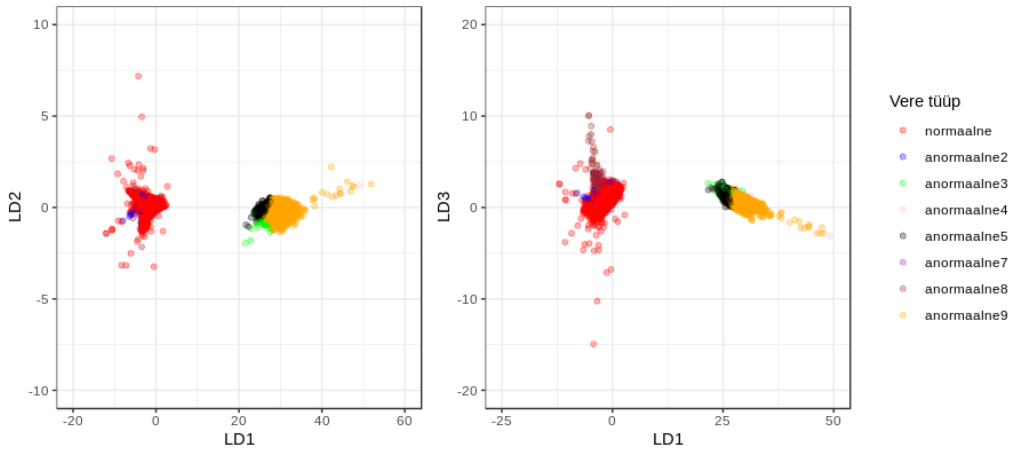
LDA leiab projektsiooni, kus klastrite keskpunktid on maksimaalsel võimalikul kaugusel üksteisest, seega eristab ta ka muid jaotusi. LDA joonistel 24 ja 25 on kujutatud kaks esimest LDA komponenti. Vastavad projektsioonid on leitud treenides mudelit kogu vereanalüüside andmestiku peal, aga joonistel on välja toodud ainult angiini ja aneemia diagnoosiga patsientide vereanalüüsid.

Angiini põdejatel eristab LDA kahte gruppi (joonis 24). Punktid on värvitud Gaussi segumudeli poolt leitud klastrite järgi. Normaalse verega on samas klastris ka anormaalsed vere klastritest 2 ja 8. Esimene LDA komponent katab ära 97% ja teine komponent 2% LDA koordinaatide varieeruvusest.

Aneemia puhul on tulemused sarnased, LDA on normaalse verega lähestikku paigutanud anormaalse vere klastrid 2 ja 8. Esimene LDA komponent katab ära samuti enamiku koordinaatide varieeruvusest - 97% ja teine 2%. Seega saab jälle järeldada, et klastrid 0, 2 ja 8 on omavahel sarnased.



Joonis 24: LDA angiini diagnoosiga patsientidel



Joonis 25: LDA aneemia diagnoosiga patsientidel

4.4 Teadaolevad probleemid

Kasutatavas andmestikus on kõik puuduvad väärtused asendatud vastava analüüdi mediaanväärtusega. Tegelikkuses võib aga puuduv väärtus olla signaal, et arst jättis selle analüüsi meelega tellimata, sest antud diagnoosi puhul oli see üleliigne. Seega tuleks puuduvaid analüüte samuti analüüsis arvesse võtta. Teinegi seletus puuduvatele väärtustele on see, et andmed pole täielikult puhastatud ja mõned väärtused lähevad kaduma. Puuduvatest väärtustest tingitud signaale arvesse võttes võiksid kõik tulemused tulla täpsemad.

Trajektooride puhul on meil teada vaid vereanalüüside tegemise kuupäevad. Tegelik haigestumise kuupäev on aga teadmata, sest mõnikord ei pöördu patsient kohe haigestudes arsti juurde, vaid ootab veidi või haigestub patsient nädalavahetusel ja peab perearsti juurde pöördumisega ootama. Seega ei saa peatükis 4.2 trajektooride alguspunkti väga täpselt määratleda. Kui täpsed haiguse algused oleks teada, siis võiksid välja joonistuda ka konkreetsemad trajektooride mustrid.

Kokkuvõte

Käesoleva töö eesmärk oli anda ülevaade eestlaste vereanalüüside tulemustes ja nende muutumisest ajas. Selleks tehti TEHIK-u vereanalüüside andmetele erinevaid esmaseid analüüse ja vaadeldi verenäitajaid nii kogu populatsioonis kui ka vanusegrupiti ja sugude kaupa.

Analüüsimaks tervete ja haigete patsientide verepilti viidi töös läbi vere klasterdamine normaalseks ja anormaalseks vereks. Klasterdamise kontrollimiseks leiti normaalse vere kõigile analüütidele referentsvahemikud ja võrreldi neid Tartu Ülikooli Kliinikumi referentsvahemikega. Ühtegi vastuolu vahemikes ei ilmnenud. Lisaks kontrolliti klasterdust ka diagnooside esinemissageduste šansside suhte leidmise abil. Ka selle puhul kinnitasid tulemused, et klasterdus tundus vastavat tegelikkusele.

Töös tegeldi ka anormaalse vere klasterdamisega. Igas anormaalse vere klastris oli vähemalt ühe analüüdi väärtus paigast ära ja domineerisid üle 60 aastased patsiendid. Anormaalse vere klasterduse korrektsuse kontrollimiseks läheks tarvis põhjalikumaid meditsiinilisi teadmisi.

Lähemalt uuriti angiini ja aneemia diagnoosiga patsientide vere muutumise trajektoore, mille puhul paistis välja, et mõlema diagnoosi korral veri püsis enamjaolt samas klastris. Angiini puhul muudeti aeg järjestustunnuseks viiepäevaste vahemikega, mis tõi välja, et peamiselt tervenesid patsiendid, kelle veri oli eelnevalt anormaalses klastris 2 või 8. Aneemia puhul võeti ajavahemikuks üks kuu, kus ilmnes, et esimestel kuudel on verepilt väga muutlik ja jällegi paranevad patsiendid, kelle veri on eelnevalt olnud klastris 2 või 8. Klastrite omavaheliste seoste uurimiseks käsitleti ka klastrisse kuulumise tõenäosusi. Sealt tuli samuti esile seosed klastrite 0, 2 ja 8 vahel.

Klastrite visuaalse eristumise nägemiseks kasutati töös lineaarset diskriminantanalüüsi ja peakomponentide analüüsi. PCA puhul ilmnes, et anormaalse verel on palju eristuvaid väärtusi. LDA puhul sattusid joonisel klastrid 0, 2 ja 8 kokku. Jällegi tõi see kinnitust nende klastrite sarnasusest.

Kokkuvõttes võib järeldada, et katsetatud meetodid toimisid. Edaspidi tuleks teha süsteemsemat analüüsi konkreetsete haigustega, võtta arvesse puuduvaid väärtusi, täpsemalt defineerida trajektooride alguspunktid ja teha rohkem koostööd meditsiini haridusega inimestega.

Viited

[1] P,ruul K., Vainsalu, E., Keidong, L., Asser, M. (2017). *E-tervise teenuste hetkeseis – andmehõive ja aegunud dokumendistandardite kasutamisel tekkivad probleemid.* https://intra.tai.ee/images/5_tervishoiustatistika2017_TEHIK_M.Asser.pdf (03.05.2019).

[2] World Health Organization (2016). *International Statistical Classification of Diseases and Related Health Problems.* Fifth edition. France: World Health Organization. https://icd.who.int/browse10/Content/statichtml/ICD10Volume2_en_2016.pdf

[3] *Sotsiaalministeerium.* http://rhk.sm.ee/ (05.05.2019).

[4] Traat, I. (2016). *Mitmemõõtmelineanalüüs.* Loengukonspekt.

[5] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis.* Third Edition. New Jersey: Wiley.

[6] Joliffe, I. T. (2002). *Principal Component Analysis.* Second Edition. New York: Springer.

[7] Rencher, A.C., Christensen, W.F. (2012). *Methods of Multivariate Analysis.* Third Edition. New Jersey: Wiley.

[8] Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E. *Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models.* https://journal.r-project.org/archive/2016/RJ-2016-021/RJ-2016-021.pdf (14.04.2019).

[9] Bishop, C.M (2006). *Pattern Recognition and Machine Learning.* Springer.

[10] McNicholas, P. D., Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing, 18(3)*, 285-296. doi: 10.1007/s11222-008-9056-0

[11] Fraley, C., Raftery, A. E. (1998). *How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis 1.* https://www.stat.washington.edu/raftery/Research/PDF/fraley1998.pdf (16.04.2019).

[12] Aus A., (2017). *SA Tartu ülikooli kliinikumi ühendlabori kvaliteedikäsiraamat.* https://www.kliinikum.ee/yhendlabor/pildid/dokumendid/he%20referentsvrtused%20vers%2011%2005.07.2017.pdf (29.04.2019).

[13] Meilivestlus perearst Evelin Raiega (19.04.2019).

[14] *Tonsillitis: Overview.* (2013). https://www.ncbi.nlm.nih.gov/books/NBK401249/ (01.05.2019).

Lisad

Lisa 1. RHK-10 koodid ja nende tähendused

Tabel 3: RHK-10 (Rahvusvaheline Haiguste Klassifikatsioon)

Peatükk	Kood	Nimetus
I	A00-B99	Teatavad nakkus-ja parasiithaigused
II	C00-D48	Kasvajad
III	D50-D89	Vere-ja vereloomeelundite haigused ning teatavad immuunmehhanismidega seotud haigusseisundid
IV	E00-E90	Sisesekretsiooni-, toitumis-ja ainevahetushaigused
V	F00-F99	Psüühika- ja käitumishäired
VI	G00-G99	Närvisüsteemihaigused
VII	H00-H59	Silma-ja silmamanuste haigused
VIII	H60-H95	Kõrva- ja nibujätkehaigused
IX	I00-I99	Vereringeelundite haigused
X	J00-J99	Hingamiselundite haigused
XI	K00-K93	Seedeelundite haigused
XII	L00-L99	Naha- ja nahaaluskoe haigused
XIII	M00-M99	Lihasluukonna ja sidekoehaigused
XIV	N00-N99	Kuse-suguelundite haigused
XV	O00-O99	Rasedus, sünnitus ja sünnitusjärgne periood
XVI	P00-P96	Perinataal- e sünniperioodis tekkivad teatavad seisundid
XVII	Q00-Q99	Kaasasündinud väärarendid, deformatsioonid ja kromosoomianomaaliad
XVIII	R00-R99	Mujal klassifitseerimata sümptomid, tunnused ja kliiniliste ning laboratoorsete leidude hälbed
XIX	S00-T98	Vigastused, mürgistused ja teatavad muud välispõhjuste toime tagajärjed
XX	V01-Y98	Haigestumise ja surma välispõhjused
XXI	U00-U99	Koodid spetsiifiliste eesmärkide jaoks
XXII	Z00-Z99	Terviseseisundit mõjustavad tegurid ja kontaktid terviseteenistusega

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Anne Ott,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Eestlaste verepildi kirjeldav analüüs", mille juhendaja on Sven Laur, 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.

Anne Ott
12.06.2019