Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Pirge Kaasik

Mitmemõõtmeliste andmete visualiseerimine meetodi *t*-SNE abil

Matemaatilise statistika eriala
Bakalaureusetöö (9 EAP)

Juhendaja: vanemteadur Kristi Kuljus

Tartu 2017

Mitmemõõtmeliste andmete visualiseerimine meetodi *t*-SNE abil

Bakalaureusetöö

Pirge Kaasik

Lühikokkuvõte. Suurte andmehulkade analüüsimise üks tähtis osa on andmete visualiseerimine. Rohkete tunnuste korral on seda tavaliste visualiseerimistehnikatega keeruline teha. Käesolevas töös on kirjeldatud mitmemõõtmeliste andmete visualiseerimismeetodit *t*-SNE, mis põhineb kõrgemamõõtmeliste andmete projekteerimisel kahe- või kolmemõõtmelisse ruumi. Meetod *t*-SNE on eelkõige andmete visualiseerimismeetod, mis püüab säilitada andmete lokaalset struktuuri vaatlustevaheliste sarnasuste abil. Meetodit on rakendatud kahe inimese pea kompuutertomograafia ja magnetresonantstomograafia mõõtmistele. Visualiseerimise tulemusena on näha, missugustes pea osades millised koed domineerivad ja millised puuduvad. Taoline informatsioon võib tulla kasuks erinevates modelleerimisülesannetes, kui on oluline andmetest eraldada informatiivsed ja vähem informatiivsed osad.

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

Märksõnad: visualiseerimine, kõrgdimensionaalsed andmed, tõenäosusjaotused, hajuvusdiagramm, kompuutertomograafia, magnetresonantstomograafia

Visualisation of multivariate data using *t*-SNE method

Bachelor's thesis

Pirge Kaasik

Abstract. Data visualisation has an important role in data analysis. There are not many good visualisation techniques for high-dimensional data. This bachelor's thesis provides a description of a method called *t*-SNE, which projects high-dimensional data into two- or three-dimensional data. Method *t*-SNE is first and foremost a visualisation method that tries to preserve local structure of the data by using pairwise similarities. The method is applied to computed tomography and magnetic resonance measurements of human heads. The results show what tissue types dominate in different parts of the head or what types are missing. Such information can be useful in different modelling tasks for selecting most informative parts of data.

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

Keywords: visualisation, high-dimensional data, probability distributions, scatterplot, computed tomography, magnetic resonance imaging

Sisukord

Sissejuhatus

Üha enam otsitakse keerulistele küsimustele vastuseid suurtelt andmehulkadelt. Mida aeg edasi, seda mahukamaks ja komplitseeritumaks andmed muutuvad. Andmete keerukus tuleneb nii nende mitmemõõtmelisusest ehk tunnuste rohkusest, vaatluste arvust kui ka tunnustevaheliste sõltuvuste struktuurist. Enne andmete analüüsimist tuleks mõista andmete iseloomu ja struktuuri. Selleks kasutatakse andmete visualiseerimist, mille eesmärk on esialgse ettekujutuse ja ülevaate andmine andmestikus olevatest tunnustest. Seeläbi on võimalik näha andmete struktuuri ja võimalikke tunnustevahelisi seoseid. Andmete illustreerimiseks kasutatakse enamasti histo-, hajuvus- ja karpdiagrammi. Need võimaldavad korraga visualiseerida aga vähe tunnuseid, mistõttu on raske ettekujutust saada tunnuste omavahelistest seostest. Seepärast on loodud erinevaid mitmemõõtmelistele andmetele mõeldud visualiseerimistehnikaid, näiteks paralleelkoordinaatide meetod (*parallel coordinates*), radiaaldiagramm (*star plot*) ja Chernoffi näod (*Chernoff faces*). Need visualiseerimistehnikad ei anna aga piisavat ülevaadet uuritavatest andmetest.

Üheks mitmemõõtmeliste andmete visualiseerimise viisiks on andmete projekteerimine madalamamõõtmelisse ruumi, eesmärgiga säilitada mitmemõõtmeliste andmete struktuur madalamamõõtmelises ruumis. Käesoleva bakalaureusetöö esimene eesmärk on kirjeldada mitmemõõtmeliste andmete visualiseerimise meetodit t-SNE (*t-distributed Stochastic Neighbor Embedding*). Meetodi t-SNE abil on võimalik projekteerida mitmemõõtmelised andmed kahe- või kolmemõõtmelisse ruumi. Antud töös käsitleme vaid kahemõõtmelist ruumi ning seega on vaatluste projektsioonid kujutatavad hajuvusdiagrammi abil. Töö teine eesmärk on rakendada meetodit t-SNE kahe inimese pea kompuutertomograafia (KT) ja magnetresonantstomograafia (MRT) mõõtmistele ning seeläbi mõista, kas ja kui hästi meetod selliste andmete korral töötab.

1 Meetod *t*-SNE

Meetod t-SNE (van der Maaten ja Hinton, 2008) on meetodi SNE (Hinton ja Roweis, 2002) edasiarendus. Mõlemad meetodid kasutavad mitmemõõtmeliste andmete projekteerimisel madalamamõõtmelisse ruumi punktide omavahelisi sarnasusi. SNE leiab need sarnasused mõlemas ruumis normaaljaotuse abil. Normaaljaotuse kasutamine sarnasuste leidmiseks on loomulik, kuna tihedusfunktsiooni väärtused kahanevad punktidevahelise kauguse kasvades. Meetodi kirjeldusest selgub, et see võimaldab iga vaatluse korral dispersioonide abil arvesse võtta selle vaatluse naabruskonda. Ka meetod t-SNE kasutab kõrgemamõõtmelises ruumis sarnasuste leidmiseks normaaljaotust, ent madalamamõõtmelises ruumis kasutatakse sarnasuste arvutamiseks t-jaotust. Sellest tuleneb ka meetodi nimi t-SNE. Üleminek t-jaotusele aitab lahendada meetodi SNE korral tekkivaid arvutuslikke probleeme. Veelgi olulisem on aga, et normaaljaotusest raskemate sabadega t-jaotus võimaldab vähendada niinimetatud ülerahvastatuse probleemi (*crowding problem*) mõju, mis on tavaline nähtus punktide projekteerimisel kõrgemamõõtmelisest madalamasse. Probleem seisneb selles, et kõrgemamõõtmeliste vaatluste paigutamiseks on madalamamõõtmelises ruumis dimensioonide erinevuse tõttu hulga vähem ruumi. Kuigi meetodit t-SNE võib vaadelda ka kui dimensioonide vähendamise meetodit, on see eelkõige visualiseerimismeetod.

1.1 Meetodi kirjeldus

Meetodi t-SNE kirjeldamisel kasutame artiklit "Visualizing data using t-SNE" (van der Maaten ja Hinton, 2008). Vaatleme p tunnusega juhuslikku vektorit $X = (X_1, \ldots, X_p)^T$. Olgu vaatlused $x_1, \ldots, x_n$ selle juhusliku vektori realisatsioonid, seega asuvad need vaatlused ruumis $\mathbb{R}^p$. Meetodi t-SNE abil teisendatakse p-mõõtmelised vaatlused punktideks $y_1, \ldots, y_n$ ruumis $\mathbb{R}^2$ nii, et säiliksid esialgsete vaatluste vahelised sarnasused. Vaatluste sarnasuste mõõtmiseks defineeritakse kummaski ruumis sarnasusmõõdud. Kuna sarnasusmõõdud on defineeritud nii, et tegemist on tõenäosusmõõtudega, saab mõõtudevahelise kauguse leidmiseks kasutada Kullback-Leibleri kaugust. Meetod t-SNE leiab punktid $y_1, \ldots, y_n$ nii, et vastav Kullback-Leibleri kaugus oleks minimaalne.

Statistikas kasutatakse vaatlustevaheliste sarnasuste kirjeldamiseks sarnasusmaatriksit (Härdle ja Simar, 2012, lk 387). Sarnasusmaatriksi elemendid võivad kirjeldada vaatlustevahelisi kauguseid või sarnasusi. Juhul kui tegemist on kaugustega, siis mida suurem on vaatlustevaheline kaugus, seda erinevamad on vaatlused. Kui sarnasusmaatriksi elemendid on vaatlustevahelised sarnasused, siis kehtib seos – mida suurem on vaatlustevaheline sarnasus, seda sarnasemad on vaatlused. Meetodi t-SNE korral on vaatluse all sarnasus-

maatriks, mille elemendid on määratud sarnasusmõõduga.

Definitsioon 1. *Vaatluste $x_1, \ldots, x_n$ sarnasusmaatriks $D : n \times n$ on defineeritud järgmiselt:*

$$D = \begin{pmatrix} d_{11} & d_{12} & \ldots & d_{1n} \\ d_{21} & d_{22} & \ldots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \ldots & d_{nn} \end{pmatrix},$$

kus d_{ij} tähistab vaatluste x_i ja x_j vahelist sarnasust. Sarnasused iseendaga d_{ii} väärtustatakse arvuga 0.

Mida suurem on sarnasusmõõt d_{ij}, seda sarnasemad on vaatlused x_i ja x_j. Meetodi t-SNE korral kasutatakse kahte seesugust sarnasusmaatriksit. Olgu $P = (p_{ij})$ vaatluste $x_1, \ldots, x_n$ ja $Q = (q_{ij})$ punktide $y_1, \ldots, y_n$ sarnasusmaatriksid. Sarnasuste p_{ij} arvutamiseks kasutatakse mitmemõõtmelise normaaljaotuse tihedusfunktsiooni ja sarnasuste q_{ij} leidmiseks standardse t-jaotuse tuuma.

1. Mitmemõõtmelise normaaljaotusega juhusliku vektori $X \sim \mathcal{N}(\mu, \Sigma)$ tihedusfunktsioon avaldub kujul

   $$f(x|\mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\Big[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\Big],$$

   kus μ ja Σ on vastavalt juhusliku vektori X keskväärtusvektor ja kovariatsioonimaatriks ehk $\mu = EX$ ja $\Sigma = cov(X)$. Erijuhul $\Sigma = \sigma^2 I_p$, kus $\sigma^2 > 0$, saame tiheduse kujul

   $$f(x|\mu, \sigma^2 I_p) = \frac{1}{(2\pi\sigma^2)^{p/2}} \exp[-\|x-\mu\|^2/(2\sigma^2)],$$

   kus $\|x - \mu\|$ tähistab vaatluse x ja μ vahelist eukleidilist kaugust.

2. Vabadusastmete arvuga 1 standardse t-jaotuse tihedusfunktsioon avaldub kujul

   $$f(t) = \frac{1}{\sqrt{\pi}}(1+t^2)^{-1}$$

   ning t-jaotuse tuum $K_t(x, y)$, kus $x, y \in \mathbb{R}^p$, on defineeritud järgmiselt:

   $$K_t(x, y) = (1 + \|x-y\|^2)^{-1}.$$

Definitsioon 2. *Olgu $x_1, \ldots, x_n$ vaatlused ruumis $\mathbb{R}^p$. Siis vaatluste x_i ja x_j vaheline sarnasusmõõt p_{ij} on defineeritud järgmiselt:*

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n},$$

kus $p_{j|i}$ on tõenäosus, et vaatlus x_i valib oma naabriks vaatluse x_j.

Tõenäosus $p_{j|i}$ avaldub mitmemõõtmelise normaaljaotuse $\mathcal{N}(x_i, {\sigma_i}^2 I_p)$ tihedusfunktsiooni $f(x|x_i, {\sigma_i}^2 I_p)$ kaudu järgmiselt:

$$p_{j|i} = \frac{f(x_j|x_i, {\sigma_i}^2 I_p)}{\sum_{k\neq i} f(x_k|x_i, {\sigma_i}^2 I_p)} = \frac{\exp[-\|x_i - x_j\|^2/(2{\sigma_i}^2)]}{\sum_{k\neq i} \exp[-\|x_i - x_k\|^2/(2{\sigma_i}^2)]},$$

kusjuures tõenäosus $p_{i|i} = 0$ ehk vaatlus x_i ei saa olla iseenda naaber.

Avaldisest on näha, et $\sum_{j=1}^{n} p_{j|i} = 1$, millest järeldub, et $\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij} = 1$. Sarnasuste p_{ij} arvutamiseks on vaja määrata dispersioon σ_i^2, mis võib sõltuda vaatlusest x_i. Seda, kuidas määratakse σ_i^2, selgitame hiljem. Nii nagu ruumis $\mathbb{R}^p$ leitakse suurused p_{ij}, arvutatakse ruumis $\mathbb{R}^2$ vastavad suurused q_{ij}, mis kirjeldavad punktide y_i ja y_j vahelist sarnasust.

Definitsioon 3. *Olgu $y_1, \ldots, y_n$ punktid ruumis $\mathbb{R}^2$. Punktide y_i ja y_j vaheline sarnasusmõõt q_{ij} arvutatakse t-jaotuse tuuma abil järgmiselt:*

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}},$$

kus $\|y_i - y_j\|$ tähistab punktide y_i ja y_j vahelist eukleidilist kaugust.

Paneme tähele, et $p_{ij} = p_{ji}$ ja $q_{ij} = q_{ji}$ ehk mõlemad sarnasusmõõdud on sümmeetrilised. Lisaks sellele on maatriksid P ja Q tõenäosusmõõdud, see tähendab, et mõlema maatriksi elemendid summeeruvad arvuks 1. Seega võib vaadelda sarnasusmaatrikseid P ja Q kui tõenäosusjaotusi.

Meetodi t-SNE idee on, et kui punktid y_i ja y_j modelleerivad vaatluste x_i ja x_j vahelist sarnasust korrektselt, siis on ka sarnasusmõõdud p_{ij} ja q_{ij} lähedased. Tõenäosusjaotuste P ja Q vahelise kauguse leidmisel kasutatakse Kullback-Leibleri kaugust (Lember, 2013; Bishop, 2009, lk 55). Kullback-Leibleri kaugus põhineb ideel, et tegelikku jaotust P lähendatakse jaotuse Q abil.

Definitsioon 4. *Olgu P ja Q kaks diskreetset tõenäosusjaotust hulgal $\chi = \{x_1, x_2, \ldots\}$. Tähistagu $p_i = P(x_i)$ ja $q_i = Q(x_i)$. Kullback-Leibleri kaugus jaotuste P ja Q vahel on defineeritud järgmiselt:*

$$KL(P\|Q) = \sum_i p_i \log \frac{p_i}{q_i}.$$

Seejuures defineeritakse $0 \log \frac{0}{q_i} = 0$, kui $q_i \geqslant 0$, ja $p_i \log \frac{p_i}{0} = \infty$, kui $p_i > 0$.

Punktid $y_1, \ldots, y_n$ leitakse nii, et P ja Q vaheline Kullback-Leibleri kaugus oleks minimaalne. Seega defineeritakse kaofunktsioon C sarnasusmaatriksite P ja Q korral nende Kullback-Leibleri kauguse kaudu järgmiselt:

$$C(y_1, \ldots, y_n) = KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

Kullback-Leibleri kauguse omadusteks on mittenegatiivsus ehk $KL(P\|Q) \geqslant 0$, ja ebasümmeetrilisus ehk $KL(P\|Q) \neq KL(Q\|P)$ (Bishop, 2009, lk 55). Kusjuures $KL(P\|Q) = 0$ siis ja ainult siis, kui $P = Q$. Kuna Kullback-Leibleri kaugus ei ole sümmeetriline, saavad erinevat tüüpi vead punktide projekteerimisel madalamamõõtmelisse ruumi erinevad kaalud. Üldiselt saame eristada kaht tüüpi vigu:

1) punktid x_i ja x_j on sarnased, aga y_i ja y_j ei ole ehk p_{ij} on suur, ent q_{ij} on väike;

2) punktid x_i ja x_j on erinevad, aga y_i ja y_j on sarnased ehk p_{ij} on väike, ent q_{ij} on suur.

Esimest tüüpi viga saab kaofunktsioonis C palju suurema kaalu kui teist tüüpi viga. Seetõttu püütakse vältida eelkõige esimest tüüpi vigu, mis tähendab, et meetodi t-SNE kaofunktsioon on suunatud andmestiku lokaalse struktuuri säilitamisele. Eesmärk on muuta kaofunktsiooni C väärtus nii väikeseks kui võimalik. Kaofunktsiooni C minimeerimiseks kasutatakse gradiendi kiirema languse meetodit (*gradient descent*). Gradiendi kiirema languse meetod põhineb ideel, et gradient on 0, kui ta asub funktsiooni lokaalses miinimumis. Selle meetodi käigus alustatakse mingist punktist ning liigutakse sammhaaval mööda funktsiooni edasi ning otsitakse kohta, kus funktsiooni gradient on 0. Sammude abil konstrueeritakse lähendite jada, mis läheneb lokaalsele miinimumile.

Tõenäosuste $p_{j|i}$ arvutamiseks on vaja määrata dispersioon ${\sigma_i}^2$. Tuletame meelde, et $p_{j|i}$ tähistab tõenäosust, et vaatlus x_i valib oma naabriks vaatluse x_j. Seejuures ei pruugi $p_{j|i}$ võrduda tõenäosusega, et vaatlus x_j valib oma naabriks vaatluse x_i ehk suurusega $p_{i|j}$. On loomulik, et $p_{j|i}$ ja $p_{i|j}$ võivad olla üsna erinevad, sest x_i ja x_j võivad paikneda nii, et nende ümber on vaatlusi erineva tihedusega. Seetõttu määrataksegi vaatlusega x_i seotud normaaljaotuse dispersioonid ${\sigma_i}^2$ vaatluse x_i naabrite arvu järgi. Dispersiooni väärtus sõltub x_i asukohast teiste punktide suhtes ehk x_i naabruskonnast. Ruumi osas, kus vaatlused paiknevad üsna tihedalt, on sobivam kasutada madalamat väärtust ja vastupidi. Sobiva ${\sigma_i}^2$ määramiseks kasutatakse entroopia (Khinchin, 1957, lk 2–4) ja perplekssuse mõistet.

Definitsioon 5. *Olgu X diskreetne juhuslik suurus jaotusega P ning olgu $\chi = \{x_1, \ldots, x_n\}$ juhusliku suuruse X võimalikud väärtused. Tähistagu $p_k = P(x_k)$. Jaotuse P entroopia on defineeritud võrdusega*

$$H(P) = H(p_1, ..., p_n) = -\sum_{k=1}^{n} p_k \log p_k.$$

Kui $p_k = 0$, siis defineeritakse $p_k \log p_k = 0$.

Entroopia on suurus, mis mõõdab sündmuste ettearvamatust. Olgu tegemist lõpliku arvu sündmustega $A_1, ..., A_n$, mille esinemistõenäosused on vastavalt $p_1, ..., p_n$. On selge, et

sündmuse ettearvamatus sõltub tema tõenäosusest. Kui näiteks $p_i = 1$ mingi i korral, siis on juhuslikkuse määr kõige väiksem ehk $H(p_1, ..., p_n) = 0$. Juhuslikkuse määr on 0, kuna on ette teada, et toimub sündmus A_i. Kui aga esinemistõenäosused on võrdsed, on võimatu ette teada, missugune sündmus toimub. Seetõttu on juhuslikkuse määr sel juhul kõrgeim ehk $H(p_1, ..., p_n) = \log n$.

Meetodi t-SNE korral leitakse iga vaatluse x_i jaoks ${\sigma_i}^2$ etteantud naabrite arvu ehk perplekssuse kaudu. Tähistagu P_i tõenäosuste $p_{j|i}$ poolt määratud jaotust, kus $j = 1, \ldots, n$ ja $j \neq i$. Jaotuse P_i entroopia on definitsiooni kohaselt

$$H(P_i) = -\sum_{j=1}^{n} p_{j|i} \log p_{j|i}.$$

Jaotus P_i sõltub parameetrist σ_i^2 ning selle jaotuse entroopia kasvab σ_i^2 kasvades. Kuna jaotus P_i ei sisalda iseenda naabriks valimise tõenäosust $p_{i|i}$, siis võrdsete tõenäosuste $p_{j|i}$ korral on $H(P_i) = \log(n-1)$. Jaotuse P_i entroopia abil leitakse iga vaatluse x_i jaoks optimaalne σ_i^2 nii, et σ_i^2 oleks vastavuses etteantud naabrite arvuga. Seega võib lihtsustatult öelda, et perplekssus on mõõt, mis näitab naabrite arvu, mille põhjal arvutatakse sarnasused p_{ij}.

Definitsioon 6. *Jaotuse P_i perplekssus on defineeritud võrdusega*

$$Perp(P_i) = e^{H(P_i)},$$

kus $H(P_i)$ on jaotuse P_i entroopia.

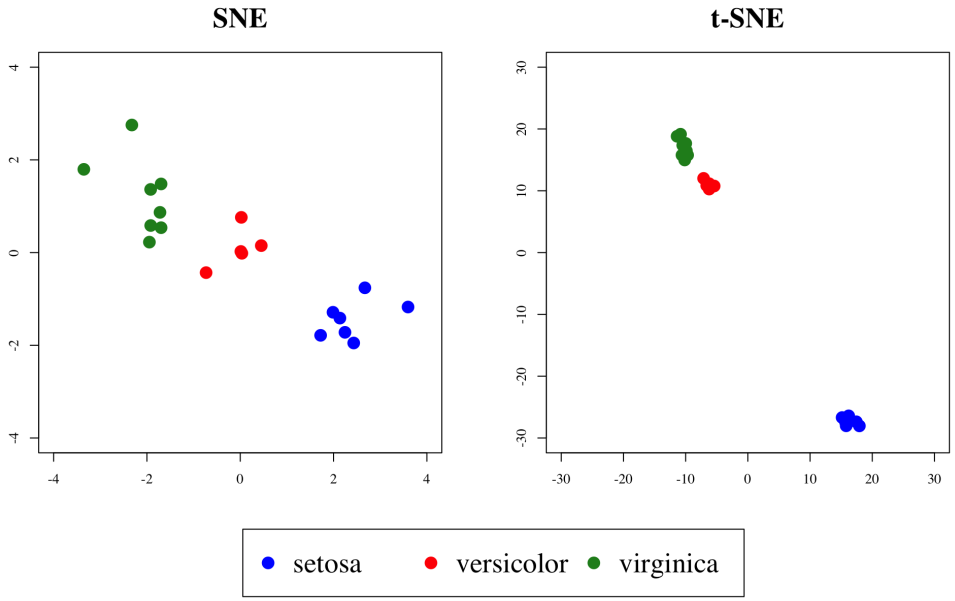
Eelnevast arutelust tuleb välja, et kui x_i naabrite valimise tõenäosusjaotus P_i on ühtlane jaotus, siis on perplekssus $n-1$. Teisel äärmuslikul juhul on vaatlusel x_i vaid üks võimalik naaber x_j mingi j korral. Sel juhul $H(P_i) = 0$ ja perplekssus on võrdne arvuga 1. Meetodi t-SNE korral soovitavad van der Maaten ja Hinton (2008) valida perplekssuse väärtused vahemikus 5–50.

Ülerahvastatuse probleem

Sobivate vaatluste projektsioonide $y_1, \ldots, y_n$ leidmisel tuleb esile niinimetatud ülerahvastatuse probleem. See tähendab, et madalamamõõtmelised punktid on paigutatud üksteisele liiga lähedale. Vaatleme näiteks 10 000 punkti kolmemõõtmelises ruumis, mis asuvad keras raadiusega 5 cm. Proovides neid punkte projekteerida kahemõõtmelises ruumis olevasse ringi, mille raadius on 5 cm, on punktide paigutamiseks oluliselt vähem ruumi. See tekitab tihedalt asetsevaid projekteeritud punkte ehk punktide ülerahvastatust. Kuna meetod t-SNE kasutab sarnasuste arvutamiseks madalamamõõtmelises ruumis raskemate sabadega t-jaotust, on ülerahvastatuse probleem meetodi t-SNE korral väiksem kui meetodi

SNE korral. Nimelt võimaldab t-jaotus punkte madalamas ruumis hajutada – üksteisest kaugemal asuvad punktid asetatakse projekteerimisel üksteisest veelgi kaugemale.

Joonis 1 kujutab meetodi SNE ja t-SNE võrdlust iirise liikide andmestiku valimil (lihtne juhuslik valik suurusega 20). Iga iirise liigi (*Iris setosa*, *Iris versicolor* ja *Iris virginica*) kohta on andmestikus 4 erinevat tunnust. Nendele andmetele on rakendatud SNE ja t-SNE meetodit samade parameetritega. Jooniselt on näha, et meetodi t-SNE korral lükatakse sarnaste vaatluste grupid kahemõõtmelises ruumis üksteisest kaugemale kui meetodi SNE korral.



Joonis 1. Meetodite SNE ja t-SNE võrdlus iirise liikide andmestiku näitel

Algoritm

Meetod t-SNE leiab vaatluste $x_1, \ldots, x_n$ jaoks sobivad madalamamõõtmelised punktid $y_1, \ldots, y_n$ itereerimise teel. Algoritm lubab kasutajal valida perplekssust ja iteratsioonide arvu. Esmalt arvutatakse algoritmis vaatlustevahelised sarnasused ehk sarnasusmaatriks P. Maatriksi arvutamisel määratakse iga vaatluse x_i korral sobiv σ_i^2 etteantud perplekssuse parameetri järgi. Seejärel seatakse punktide $y_1, \ldots, y_n$ algväärtuseks pseudojuhuslikud arvud jaotusest $\mathcal{N}(0, 10^{-4} I_p)$. Iga sammu ehk iteratsiooni käigus arvutatakse punktide $y_1, \ldots, y_n$ sarnasusmaatriks Q ja kaofunktsiooni C gradient. Gradienti ja eelnevaid $y_1, \ldots, y_n$ väärtusi kasutades leitakse uued väärtused punktidele $y_1, \ldots, y_n$. Pärast itereerimist on tulemuseks iga esialgse vaatluse x_i projektsioon y_i ruumis $\mathbb{R}^2$. Projekteeritud punktid $y_1, \ldots, y_n$ esitatakse hajuvusdiagrammi abil.

1.2 Meetodi rakendamine

Peakomponentide meetod

Juhul kui andmestikus esineb tunnuseid liiga palju, vähendatakse enne t-SNE algoritmi rakendamist esialgsete tunnuste arvu. Tunnuste arvu vähendamiseks kasutatakse peakomponentide meetodit (Härdle ja Simar, 2015, lk 319–331). Meetodi idee on moodustada esialgsete tunnuste $X_1, X_2, \ldots, X_p$ abil uued tunnused $Z_1, Z_2, \ldots, Z_p$ nii, et uute tunnuste kogudispersioon oleks sama, mis esialgsete tunnuste kogudispersioon. Uued tunnused leitakse nii, et suur osa dispersioonist oleks kirjeldatud just esimeste tunnuste poolt. Uusi tunnuseid $Z_1, Z_2, \ldots, Z_p$ nimetatakse peakomponentideks, need on esialgsete tunnuste $X_1, X_2, \ldots, X_p$ lineaarkombinatsioonid.

Olgu Σ_X juhusliku vektori X kovariatsioonimaatriks ja olgu V maatriksi Σ_X omavektorite maatriks, kusjuures omavektorid on normeeritud ehk pikkusega üks ja asuvad maatriksi veergudes, seega $V^TV = I_p$. Maatriksi Σ_X spektraallahutust kasutades saame kovariatsioonimaatriksi esitada kujul

$$\Sigma_X = V\Lambda V^T,$$

kus Λ on maatriks, mille diagonaalil asuvad kovariatsioonimaatriksi Σ_X omaväärtused. Esialgsete tunnuste $X_1, X_2, \ldots, X_p$ kogudispersioon avaldub kujul:

$$\sum_{i=1}^{p} D(X_i) = tr(\Sigma_X) = tr(\Lambda) = \sum_{i=1}^{p} \lambda_i.$$

Uued tunnused $Z_1, Z_2, \ldots, Z_p$ moodustatakse eeskirja $Z = V^T(X - \mu_X)$ abil, kus μ_X on esialgsete tunnuste keskväärtus. Uute tunnuste kovariatsioonimaatriksi Σ_Z saame arvutada järgmiselt:

$$\Sigma_Z = E[V^T(X - \mu_X)(X - \mu_X)^T V] = V^T\Sigma_X V = (V^TV)\Lambda(V^TV) = \Lambda.$$

Järelikult on esialgsete tunnuste ja uute tunnuste kogudispersioon sama. Uued tunnused $Z_1, Z_2, \ldots, Z_p$ on järjestatud dispersiooni kahanemise järgi ja nad on lineaarselt sõltumatud. Esimese peakomponendi Z_1 dispersioon on suurim ning viimase peakomponendi Z_p dispersioon on väikseim. Seega kui r esimest peakomponenti kirjeldavad ära suure osa tunnuste hajuvusest, ei kaota me informatsiooni mõttes eriti palju, kui edasises analüüsis kasutame vaid r esimest peakomponenti.

Barnes-Huti lähendus

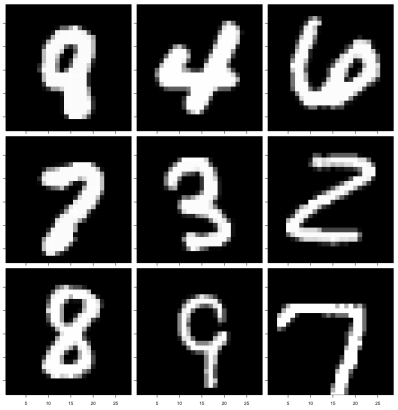
Suurte andmestike korral, kus on enam kui 10 000 vaatlust, on meetodi t-SNE kasutamine väga ajakulukas, mistõttu kasutatakse Barnes-Huti lähendust (van der Maaten, 2014).

Suurima ajakuluga on Kullback-Leibleri kauguse minimeerimisel kasutatav gradiendi kiirema languse meetod. Selle meetodi iga iteratsioon on ruutkeerukusega ehk keerukusega $O(n^2)$. See tähendab, et kui vaatluste arv on n, siis algoritm peab tegema umbes n^2 põhioperatsiooni. Barnes-Huti lähendusega on ühe iteratsiooni keerukus aga $O(n \log n)$, mis on parem kui $O(n^2)$. Seepärast kasutatakse suurte andmestike korral meetodi t-SNE kiirendamiseks Barnes-Huti lähendust. Meetodi t-SNE lähendamine põhineb kahel sammul. Esiteks leitakse tõenäosused $p_{j|i}$ vaid juhul, kui vaatlus x_j kuulub vaatluse x_i lähimate naabrite, täpsemalt $3 \cdot Perp$ lähima naabri sekka. Ülejäänud tõenäosused $p_{j|i}$ võrdsustatakse arvuga 0. Teiseks sammuks on gradiendi arvutamine, mille jaoks kasutatakse Barnes-Huti algoritmi.

Rakendustarkvaras R on meetodi t-SNE rakendamiseks funktsioon `Rtsne` (Krijthe ja van der Maaten, 2014), mis põhineb t-SNE Barnes-Huti lähendusel. Funktsiooni `Rtsne` olulisemad sisendparameetrid on peakomponentide arv `initial_dims` (vaikimisi 50), perplekssus `perp` ja `theta` (vaikimisi 0,5). Nagu eelnevalt mainitud, rakendatakse esialgsete tunnuste arvu vähendamiseks peakomponentide meetodit. Järelikult on oluline määrata uute tunnuste arv ehk mitut peakomponenti võtab edasine algoritm arvesse. Selle kaudu on võimalik määrata, kui suur osa esialgsete tunnuste kogudispersioonist algoritmi kaasatakse. Kui esialgsete tunnuste arv ei ole suur, on võimalik funktsiooni sisendparameetriga `pca` peakomponentide meetod sisse/välja lülitada (TRUE/FALSE). Teine oluline sisendparameeter, perplekssus ehk naabrite arv, määrab ära lokaalse naabruskonna suuruse. Viimaseks oluliseks parameetriks on `theta`, mis reguleerib Barnes-Huti lähenduse kiiruse ja täpsuse suhet. Parameetri `theta` väikesemad väärtused annavad täpsema lähenduse. Kui parameetri `theta` väärtus on 0, siis leitakse lähenduseta t-SNE. Olulised väljundparameetrid on kahemõõtmelised punktid $y_1, \ldots, y_n$ ja `itercost`, mis näitab Kullback-Leibleri kaugust pärast kaofunktsiooni C minimeerimist. Punktide $y_1, \ldots, y_n$ abil konstrueeritakse hajuvusdiagramm, mis kujutab esialgsete vaatluste projektsiooni madalamamõõtmelises ruumis.

1.3 Näide numbrite andmestikuga

Rakendustarkvaras R oleva funktsiooni `Rtsne` testimiseks kasutame MNIST andmestiku üht osa (LeCun et al., 1999). See sisaldab 10 000 must-valget pilti käsitsi kirjutatud numbritest $0, \ldots, 9$ (vt joonis 2). Iga pilt on kirjeldatud $28 \cdot 28 = 784$ piksli abil. Pikslid on järjestatud vektoriks, seega on pikslite arv ühtlasi ka andmestiku tunnuste arvuks ehk vaatlused asuvad ruumis $\mathbb{R}^{784}$. Sama andmestikku on ühe näiteandmestikuna kasutanud ka van der Maaten ja Hinton (2008).



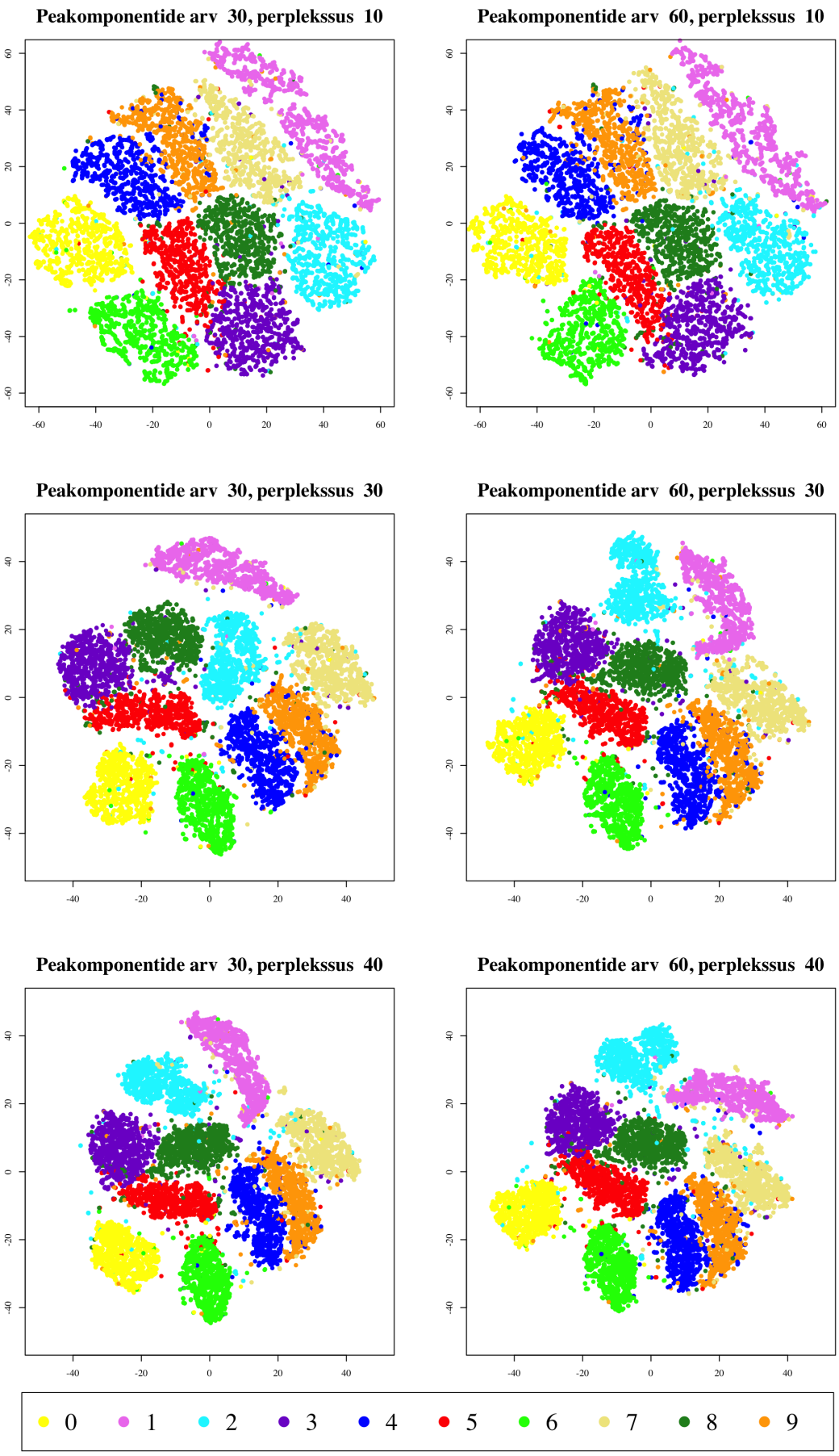
Joonis 2. MNIST andmestiku esimesed 9 vaatlust

Meetodi paremaks mõistmiseks on sobilik katsetada funktsiooni `Rtsne` erinevate sisendparameetrite väärtustega. Joonisel 3 on võrreldud peakomponentide arvu ja perplekssuse valiku mõju projektsiooni tulemusel saadud pildile. Iga pildi konstrueerimisel on kasutatud samu pseudojuhuslikke arve ning muud sisendparameetrid on jäetud vaikimisi väärtusteks.

Tabel 1. Kullback-Leibleri kaugused joonisel 3

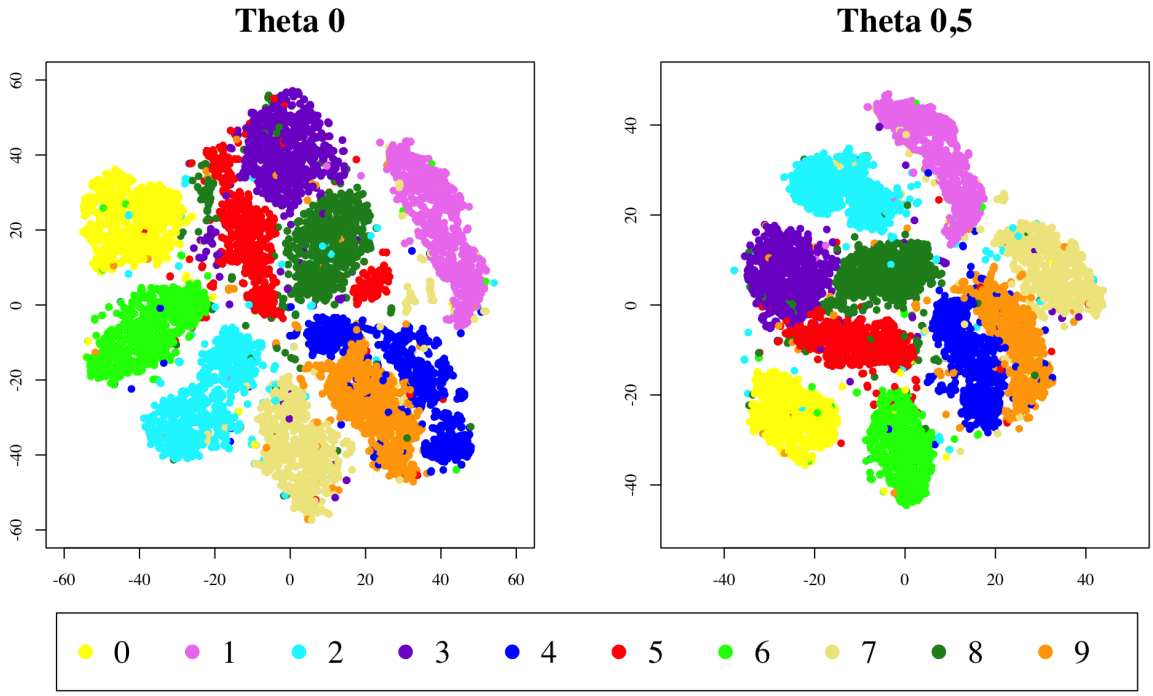
Peakomponentide arv \ Perplekssus	10	30	40
30	2,1141	1,8750	1,7996
60	2,1499	1,9170	1,8314

Esimesed 30 peakomponenti kirjeldavad ligikaudu 74% esialgsete tunnuste koguhajuvusest. Esimesed 60 peakomponenti kirjeldavad aga ligikaudu 86% esialgsete tunnuste koguhajuvusest. Tabelist 1 on näha, et perplekssuse suurenedes Kullback-Leibleri kaugus sarnasusmaatriksite P ja Q vahel väheneb. Tabeli 1 põhjal vastab kõige väiksem Kullback-Leibleri kaugus pildile, mille peakomponentide arv on 30 ja perplekssus 40. Võrreldes pilte numbrite rühmade eristamise alusel, on selge, et perplekssuse väärtuse 10 korral on numbrite rühmad üksteisest raskemini eristatavad. Antud juhul pole peakomponentide arv numbrite rühmade asetust muutnud. Küll aga on teiste perplekssuse väärtuste korral peakomponentide arv mõjutanud projektsiooni tulemusel saadud pilte. Peakomponentide arvu 60 ja perplekssuse 40 korral on numbrite rühmad 0, 1, 2 ja 6 selgelt teistest rühmadest eraldi. Sel pildil on moodustunud ka kaks numbrite rühmade kolmikut – numbrid 3, 5 ja 8 ning numbrid 4, 7 ja 9 moodustavad kaks ühtset gruppi. Visuaalselt on näha, et mõni number satub valesse rühma ehk meetodi t-SNE käigus on tehtud teist tüüpi viga. See tuleneb ilmselt sellest, et käsitsi kirjutades on mõned numbrid üksteisele väga sarnased. Näiteks on joonisel 2 esimesel pildil olev number 8 väga sarnane numbriga 9. Sellest hoolimata on meetod t-SNE suutnud numbrite rühmad hästi eraldada.



Joonis 3. MNIST andmete visualiseerimise tulemused erinevate peakomponentide arvu ja perplekssuse korral

Joonisel 4 on võrreldud visualiseerimise tulemusi kahe erineva `theta` väärtuse korral, täpsemalt 0 ja 0,5 korral. Piltide loomiseks on kasutatud samu pseudojuhuslikke arve, peakomponentide arvu 30 ja perplekssust 40. Ülejäänud sisendparameetrid on jäetud vaikimisi väärtusteks. Katsetamise käigus ilmnes, et lähenduseta t-SNE meetodi algoritmi käitusaeg on ligikaudu 20 korda suurem (ligi 44 minutit) kui `theta`=0,5 korral (ligi 2 minutit). Vaadates joonist 4 on näha, et numbrite rühmad on `theta`=0,5 korral eristunud paremini kui `theta`=0 korral. Tuleb välja, et projekteeritud pildi kvaliteet on `theta`=0,5 korral vähemalt sama hea kui `theta`=0 korral (van der Maaten, 2014). Kuna ajakulu on `theta`=0 korral oluliselt suurem ning kuna `theta`=0,5 korral pole visualiseerimise tulemus halvem, kasutame edasises töös `theta` väärtust 0,5.

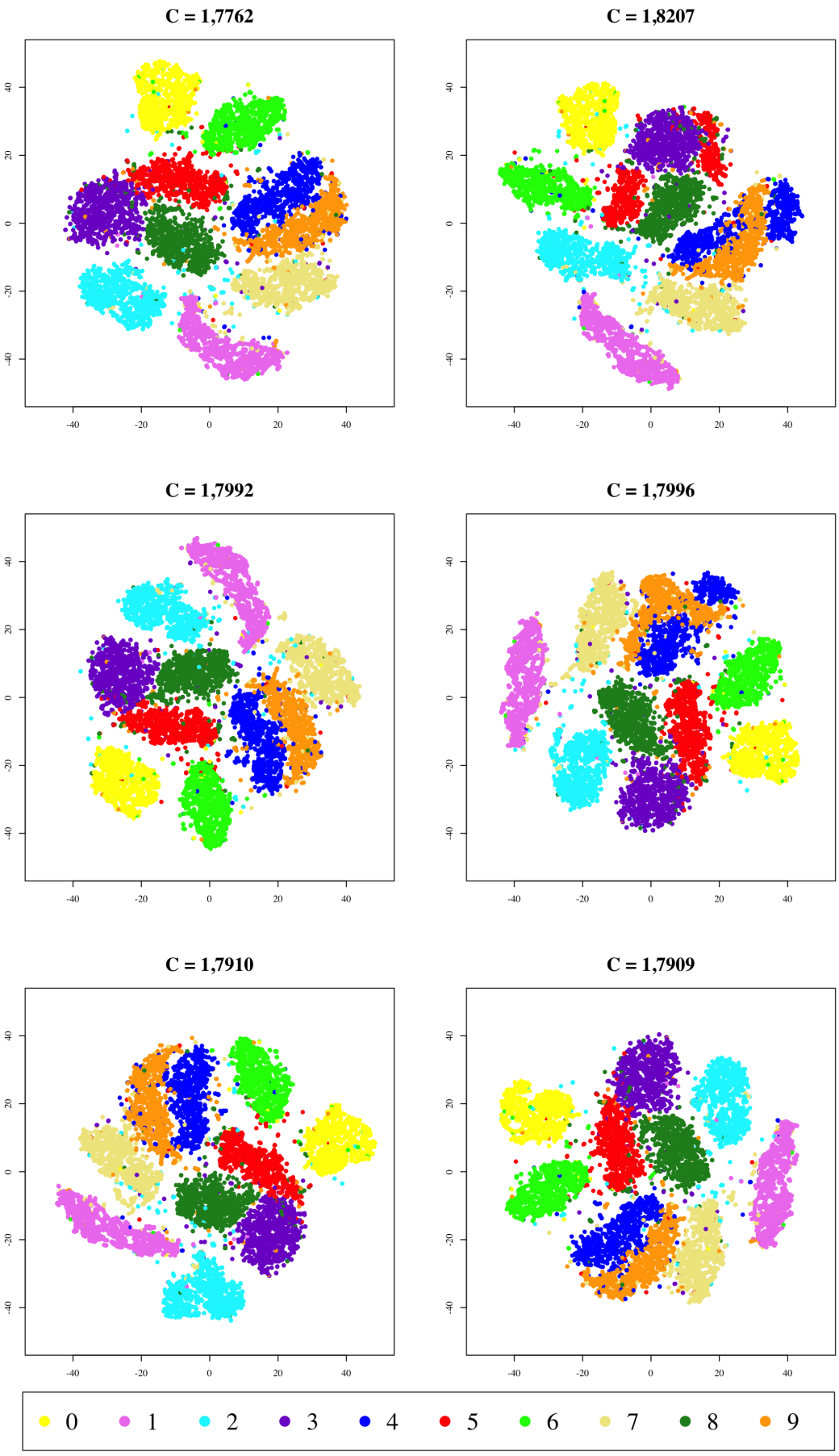


Joonis 4. MNIST andmete visualiseerimine lähenduseta (`theta`=0) ja lähendusega (`theta`=0,5) t-SNE algoritmi korral

Meetod t-SNE kasutab punktide $y_1, \ldots, y_n$ alglähendina pseudojuhuslikke arve normaaljaotusest, seega sõltuvad projekteeritud punktid nendest arvudest. Alglähendite mõju analüüsimiseks vaatleme 10 projekteeritud pilti, mis kasutavad erinevaid pseudojuhuslikke arve. Visualiseerimisel on kasutatud perplekssust 40, peakomponentide arvu 30 ja väärtust `theta`=0,5. Joonisele 5 (lk 17) on lisatud kuus saadud kahemõõtmelistest piltidest. Kõik joonisel 5 olevad pildid on asetuselt erinevad. Järelikult mõjutavad pseudojuhuslikud arvud tulemust. Nende kahemõõtmeliste piltide Kullback-Leibleri kaugused on vahemikus 1,7762–1,8207 (vt joonis 5). Joonise 5 esimeses reas on kõige väiksema Kullback-Leibleri kaugusega pilt ja kõige suurema Kullback-Leibleri kaugusega pilt. On selge, et antud

võrdluse korral on väiksema Kullback-Leibleri kaugusega pildil erinevate numbrite rühmad paremini eristunud. Kõige suurema Kullback-Leibleri kaugusega projekteeritud pildi korral on sarnased numbrid 4 ja 9 moodustanud omaette rühma. Samuti on numbrid 3 ja 5 sattunud samasse rühma. Kui aga võrrelda pilte, mille C=1,7992 ja C=1,7910, ei saa öelda, et väiksema Kullback-Leibleri kauguse korral oleks tulemus parem. Teises ja kolmandas reas on võrreldud peaaegu võrdsete Kullback-Leibleri kaugustega pilte. Vaadates teise rea pilte, mille Kullback-Leibleri kaugused on väga lähedased, on näha, et numbrite rühmad 4 ja 9 on parempoolsemal pildil segunenud. Järelikult ei saa valida sobivaimat tulemust ainult Kullback-Leibleri kauguse järgi.

Tuleb silmas pidada, et need parameetrite väärtused, mis on selle näite puhul sobivad, ei pruugi sobida teiste andmete korral. See tähendab, et iga andmestikuga tuleks läbi proovida erinevad sisendparameetrite komplektid ning leida konkreetsele andmestikule sobivad sisendparameetrite väärtused.



Joonis 5. MNIST andmete visualiseerimine kasutades erinevaid pseudojuhuslikke arve

2 Kompuuter- ja magnetresonantstomograafia mõõtmiste visualiseerimine

Meetodit t-SNE on rakendatud kahe inimese pea kompuutertomograafia (KT) ja magnetresonantstomograafia (MRT) andmetele. MRT ja KT on kaks erinevat tehnoloogiat, mis võimaldavad inimkehast saada kihilisi ja ruumilisi kujutisi. MRT põhineb magnetismil ja KT röntgenkiirgusel. Nende tehnoloogiate abil saadud pilte kasutatakse meditsiinilises diagnostikas. Vaadeldud peadest esimene (Pea 1) sisaldab naissoost isiku pea mõõtmisi ja teine (Pea 2) meessoost isiku pea mõõtmisi. Peade valik on tingitud nende käitumisest varasemas modelleerimisülesandes, mille eesmärk oli KT piltide prognoosimine MRT piltide põhjal (Kuljus et al., 2017). Osutus, et Pea 1 andis hinnatud mudeli korral halva prognoosipildi ja Pea 2 hea prognoosipildi.

Visualiseerimise eesmärk on saada parem ettekujutus andmete jaotuse kohta erinevates pea osades. Lisaks sellele soovime uurida, kas on võimalik tuvastada visuaalseid erinevusi nende kahe pea korral, mis käitusid modelleerimisülesandes nii erinevalt.

2.1 Andmete kirjeldus

Tegemist on ruumiliste andmetega, kumbki pea on kuubi sees. Suur kuup on jaotatud $192 \times 192 \times 192$ väiksemaks kuubiks ehk voksliks, kusjuures voksli küljepikkus on 1,33 mm. Seega on peade mõõtmised kirjeldatud vokslite ehk kolmemõõtmeliste elementide abil, mille asukoht kuubis on määratud koordinaatidega x, y ja z. Iga voksli jaoks on antud KT väärtus ja neli MRT mõõtmist, seega on iga voksli jaoks olemas viie tunnuse mõõtmised. Lisaks on iga voksli kohta teada binaarne tunnus `indeks`, mis eraldab suurest kuubist inimese pea ehk vaatluste piirkonna. Kui tunnuse `indeks` väärtus on 1, on tegemist vaatluste voksliga, vastasel juhul on tunnuse `indeks` väärtus 0. Neli MRT mõõtmist vastavad neljale erinevale MRT pildile, mis on saadud nelja erineva magnetvälja parameetrite komplektiga magnetkaameras. Teatavasti sõltuvad MRT signaali väärtused magnetvälja parameetritest ning parameetrite erinevad kombinatsioonid võimaldavad saada just mingi konkreetse eesmärgi jaoks optimaalseid pilte. Antud mõõtmiste korral olid parameetrid valitud nii, et saada võimalikult head informatsiooni luu kohta, kuna üldiselt on MRT piltidelt võimatu eristada luud ja õhku.

2.2 Meetodi *t*-SNE rakendamine

Pea piirkonnad on anatoomiliselt väga erinevad, mille tõttu jagame suure kuubi väikesemateks osadeks, et saada parem ülevaade kudede jaotuse kohta. Suure kuubi külgede kuueks jagamisel moodustub 216 uut kuupi, mille küljepikkus on 32 vokslit. Seega on väikeses kuubis maksimaalselt $32^3 = 32768$ vaatluste vokslit. Antud töös visualiseerime meetodi *t*-SNE abil 8 keskmist kuupi. See tähendab, et vaatluse all on tükid, mille vokslite koordinaatide x, y ja z väärtused on vahemike 65–96 ja 97–128 võimalikud kombinatsioonid. Binaarse tunnuse `indeks` abil eraldame andmestikust vaatluste vokslid. Tabelis 2 ja 3 on ära toodud mõlema pea vaatluste vokslite arvud vastavalt pea tükile, mis on nummerdatud koordinaatide kombinatsioonide järgi. Pea 2 korral on 8 kuubi vaatluste vokslite arv 262 144 ehk maksimaalne, seega on KT ja MRT mõõtmised olemas kõikide vokslite jaoks. Pea 1 korral on vaatluste vokslite arv 224 337. Hulga mõõtmisi on puudu just esialgse suure kuubi alumisel poolel ehk kui z=65–96 (vt tabel 2). Paneme tähele, et Pea 1 korral on kogu pea vaatluste vokslite arv oluliselt väiksem kui Pea 2 korral, vastavalt 1 325 922 ja 1 853 702. Vaadeldud 8 kuubi korral on Pea 1 vaatluste vokslite arv 15% väiksem kui Pea 2 korral.

Tabel 2. Pea 1 vaatluste vokslite arv

Tükk	1	2	3	4	5	6	7	8
x	65–96	65–96	65–96	65–96	97–128	97–128	97–128	97–128
y	65–96	65–96	97–128	97–128	65–96	65–96	97–128	97–128
z	65–96	97–128	65–96	97–128	65–96	97–128	65–96	97–128
Vokslite arv	21379	32768	22567	32768	24070	32768	25249	32768

Tabel 3. Pea 2 vaatluste vokslite arv

Tükk	1	2	3	4	5	6	7	8
x	65–96	65–96	65–96	65–96	97–128	97–128	97–128	97–128
y	65–96	65–96	97–128	97–128	65–96	65–96	97–128	97–128
z	65–96	97–128	65–96	97–128	65–96	97–128	65–96	97–128
Vokslite arv	32768	32768	32768	32768	32768	32768	32768	32768

Antud andmete korral on tunnuste arv väike, mistõttu pole enne meetodi *t*-SNE rakendamist tunnuste arvu vähendamine vajalik. Seega on funktsiooni `Rtsne` sisendparameeter `pca`=FALSE ehk algoritmi käigus peakomponentide meetodit ei rakendata. Kudede rühmade eristamiseks värvime projekteeritud punktid hajuvusdiagrammidel KT väärtuste

põhjal. KT väärtused antud peade andmete korral asuvad vahemikus [-1024, 2200). Võttes arvesse KT väärtusi tavalisemate kudede korral (vt lisa 1), jagame KT väärtuste piirkonna kuueks klassiks. Nii moodustuvad klassid [-1024, -800), [-800, -200), [-200, 0), [0, 70), [70, 700) ja [700, 2200). Kuna vokslid koosnevad tihti kudede segust, ei saa kõikidele klassidele vastavusse seada kindlat kudet ja seega ei saa me alati anda klassile ka konkreetset nime. Näiteks klass [-1024, -800) vastab õhule, klass [-200, 0) vastab rasvale, klass [0, 70) vastab hall- ja valgeollusele ning klass, kus KT väärtused on suuremad kui 700, vastab erinevatele luukudedele.

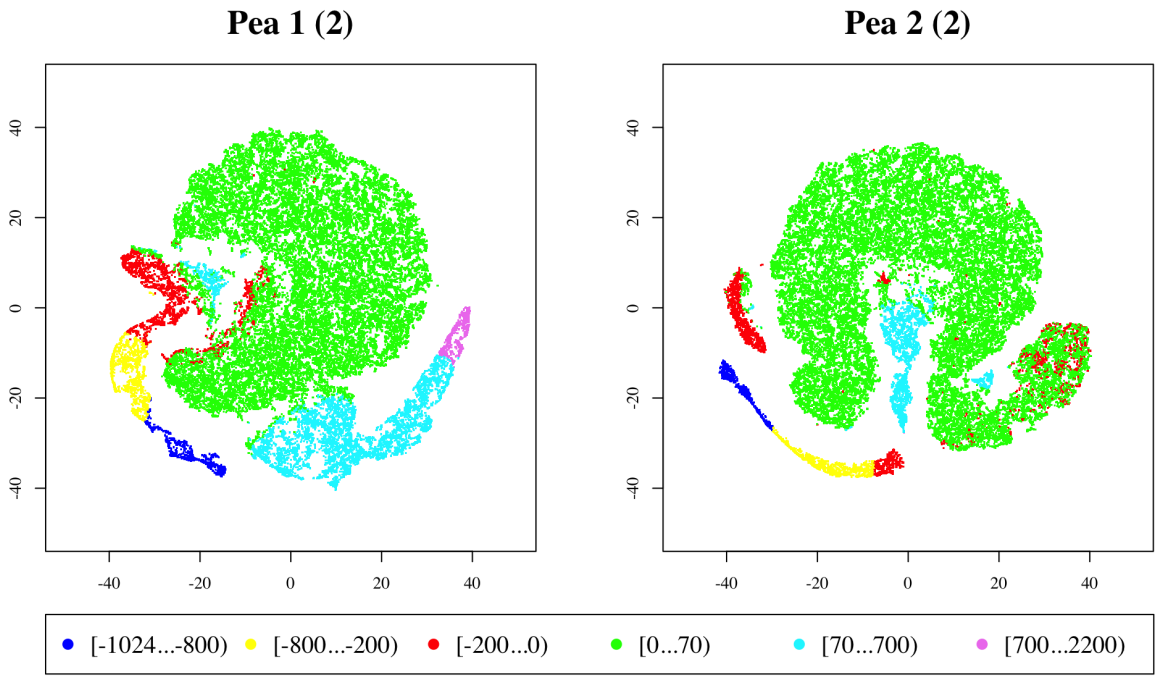
Sobiva perplekssuse väärtuse leidmiseks katsetasime nelja erinevat perplekssust: 10, 20, 40 ja 60 (vt lisa 3). Neid parameetri väärtusi proovisime Pea 1 teise tüki andmete peal. Projektsioonide piltide põhjal on antud andmete korral sobivaim perplekssuse väärtus 60, kuna vastaval hajuvusdiagrammil on kudede klassid kompaktsemalt eristatud. Sellest tulenevalt kasutame peade andmete visualiseerimiseks perplekssuse väärtust 60. Näide funktsiooni `Rtsne` kasutamise kohta on ära toodud lisas 2.

2.3 Pea andmete visualiseerimise tulemused

Antud töös analüüsime lähemalt tükkide 2 ja 7 madalamamõõtmelisi pilte Pea 1 ja Pea 2 korral. Ülejäänud tükkide projekteeritud punktide hajuvusdiagrammid on toodud lisas 4 ja 5. Kõik pildid on konstrueeritud kasutades samu pseudojuhuslikke arve ja perplekssust 60. Paneme tähele, et paarisarvuliste tükkide korral on mõlema pea vaatluste vokslite arv maksimaalne ehk puuduvaid andmeid pole. Seevastu paarituarvuliste tükkide korral on Pea 1 vaatluste vokslite arv oluliselt väiksem ehk sel peal on palju puuduvaid mõõtmisi.

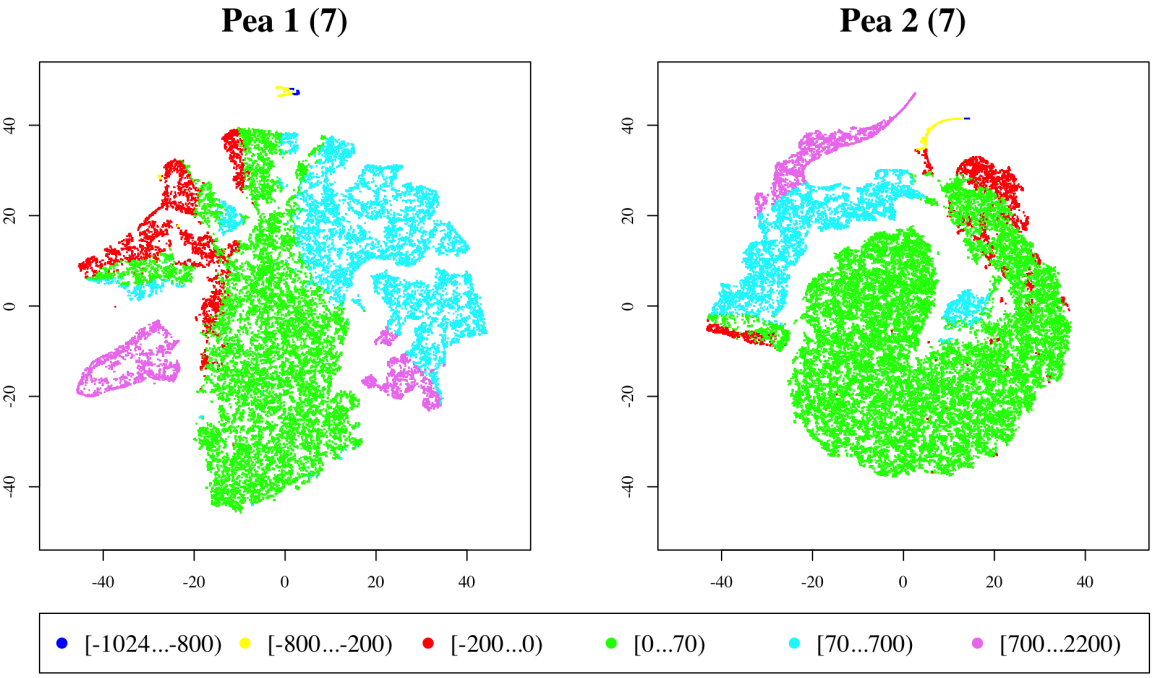
KT ja MRT mõõtmiste projekteerimine

Joonised 6 ja 7 kujutavad KT ja MRT mõõtmiste projektsioone. Joonisel 6 on näha Pea 1 ja Pea 2 teise tüki projektsioonid kahemõõtmelises ruumis. Vaadates erinevate klasside esinemist projekteeritud piltidel, on näha, et need ei ole kahe pildi korral võrdsed. Näiteks Pea 2 korral puudub pildilt luukoe klass. Klassi puudumise põhjus võib tuleneda sellest, et need kaks pead on erineva suurusega või paiknevad suure kuubi sees erinevalt, ning seega on väikeses kuubis olevad pea osad pisut erinevad. Pea 2 korral on näha, et rasvkude on mingil määral segunenud hall- ja valgeollusega. Samuti võib mõlemal pildil märgata hobuserauakujulist moodustist hall- ja valgeolluse rühmas, mis võib viidata hall- ja valgeolluse erinevusele. Projekteeritud pildid näitavad, et mõlema pea korral on klassid üsna hästi eraldunud. Joonise 6 põhjal ei saa järeldada, et need kaks pead oleksid üksteisest väga erinevad.



Joonis 6. Pea 1 ja Pea 2 tüki 2 projekteeritud punktide hajuvusdiagrammid KT ja MRT mõõtmiste korral

Joonisel 7 on toodud Pea 1 ja Pea 2 seitsmenda tüki projektsioonid kahemõõtmelises ruumis. Jooniselt on näha, et Pea 2 korral paiknevad projekteeritud andmepunktid tunduvalt kompaktsemalt kui Pea 1 korral. Üks võimalikest põhjustest on puuduvad andmed

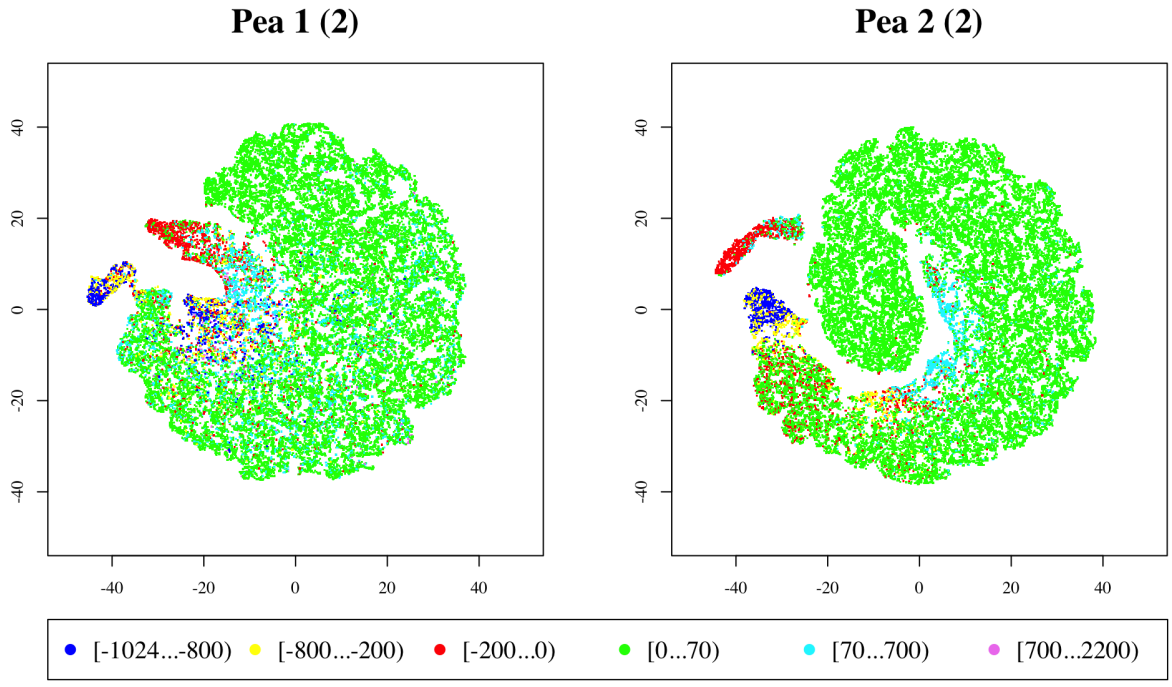


Joonis 7. Pea 1 ja Pea 2 tüki 7 projekteeritud punktide hajuvusdiagrammid KT ja MRT mõõtmiste korral

Pea 1 korral. Kuna meetod t-SNE kasutab vaatlustevahelisi sarnasusi, mis on leitud vaatluste naabrite järgi, võivad puuduvad andmed vaatlustevahelisi sarnasusi muuta. Sellest tulenevalt paigutab meetod t-SNE vaatlused madalamamõõtmelisse ruumi valesti. Pea 1 pildilt on näha, et luukudede klass paikneb kahemõõtmelisel pildil kahes rühmas. Rühm võib olla eraldatud sellepärast, et luu klassi kuuluvad erineva tihedusega luukoed. Luukudede tiheduse erinevuse tõttu võib visualiseerimismeetod need üksteisest eraldada.

MRT mõõtmiste projekteerimine

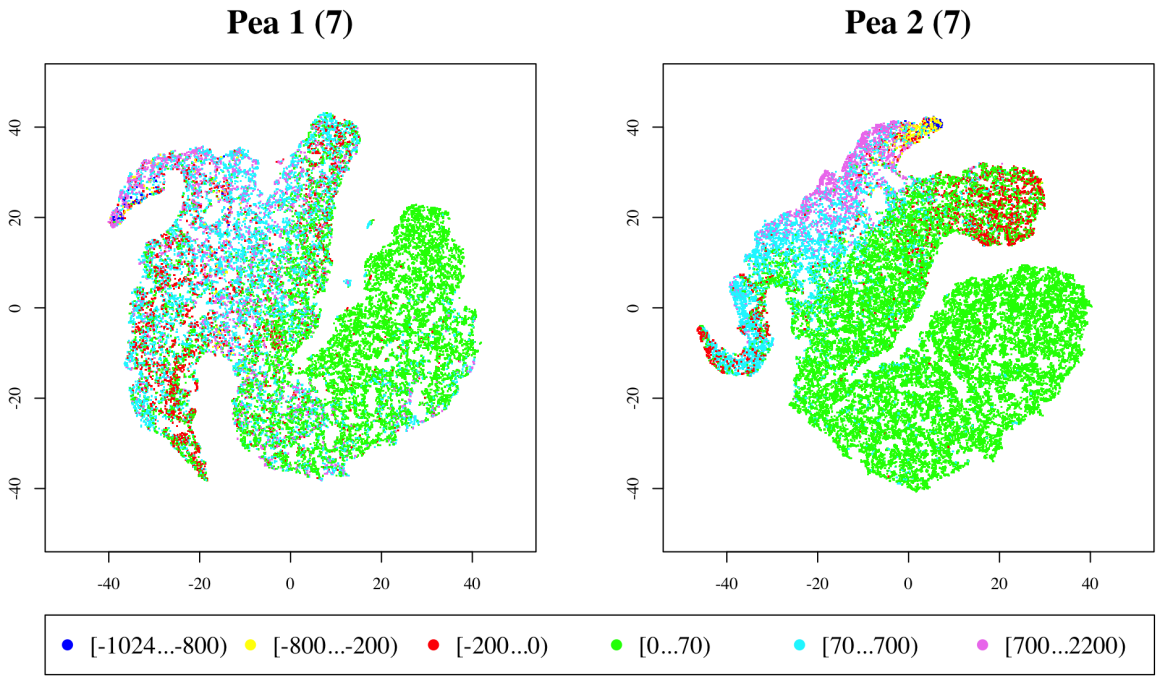
KT ja MRT mõõtmiste visualiseerimisel ei olnud näha kahe pea vahel selgeid erinevusi. Järgnevalt uurime, kas modelleerimisel ilmnenud erinevused kahe pea korral võivad olla seotud MRT andmetega. Joonised 8 ja 9 kujutavad pea andmete projektsioone, kui tunnustena on kasutatud vaid MRT mõõtmisi. Kudede klassid on endiselt defineeritud KT väärtuste põhjal. Joonisel 8 on Pea 1 ja Pea 2 MRT mõõtmiste põhjal konstrueeritud teise tüki visualisatsioonid. Meetod t-SNE koondab osaliselt vaid mõned klassid. Näiteks on mõlema pea korral madalamamõõtmelisel pildil moodustunud hall- ja valgeolluse rühm. Samuti on mingil määral koondunud ka rasva ja õhu klass. Nagu eelnevalt sai mainitud, on MRT korral raske eristada luud ja õhku. Seda on näha ka Pea 1 piltide võrdlusest joonise 6 ja joonise 8 korral – joonisel 8 on luu rühm ära kadunud. Paljud luu projekteeritud punktid on laiali paisatud, teisalt on arvatavasti osa punkte paigutatud õhu klassiga sa-



Joonis 8. Pea 1 ja Pea 2 tüki 2 projekteeritud punktide hajuvusdiagrammid MRT mõõtmiste korral

masesse rühma. Üldiselt võib öelda, et Pea 2 korral on erinevatele kudedele vastavad grupid selgemini eristatavad.

Joonis 9 kujutab Pea 1 ja Pea 2 seitsmenda tüki MRT mõõtmiste põhjal loodud madalamamõõtmelised pildid. On näha, et Pea 1 korral on erinevad koed väga halvasti eraldunud. See tähendab, et meetodi t-SNE arvates ei koonda meie poolt defineeritud kudede klassid omavahel sarnaseid punkte. Joonise 9 vasakpoolne pilt näitab selgelt, et t-SNE meetod ei säilita projekteerimisel meie poolt defineeritud lokaalset struktuuri. Pea 1 korral säilib lokaalne struktuur mingil määral ainult hall- ja valgeolluse klassi korral. Pea 2 korral on klassid oluliselt paremini koondunud kui Pea 1 korral. Põhjuseks võib olla MRT mõõtmiste halb kvaliteet Pea 1 korral, mistõttu ei ole võimalik erinevaid kudesid eristada. Kõikide paarituarvuliste tükkide korral on Pea 1 projekteeritud punktide hajuvusdiagrammidel KT väärtuste klassid segunenud ja eristamatud (vt joonis 13, lk 30).



Joonis 9. Pea 1 ja Pea 2 tüki 7 projekteeritud punktide hajuvusdiagrammid MRT mõõtmiste korral

Kokkuvõte

Meetodi t-SNE rakendamine KT ja MRT andmetele näitab, et visualiseerimise eesmärgid on täidetud. Just KT ja MRT mõõtmiste ühine projekteerimine kinnitab, et meetod t-SNE säilitab andmete lokaalse struktuuri. See tähendab, et samasse klassi kuuluvad vaatluste vokslid on üksteise lähedal ka madalamamõõtmelisel pildil. Samuti on meetod proovinud säilitada andmete globaalset struktuuri. Kõrvuti paiknevad KT väärtuste klassid asuvad ka projekteeritud pildil üksteise kõrval ja need koed, mis pole KT väärtuste

põhjal üksteisele lähedal on ka projekteeritud pildil üksteisest eraldatud. Näiteks on klass [700, 2200) alati klassi [70, 700) kõrval. Üksteisest KT väärtuste põhjal kõige kaugemal asuvad klassid õhk ja luu pole kunagi madalamamõõtmelisel pildil omavahel ühenduses. Ka teine visualiseerimise eesmärk on täidetud – nimelt on tükkide projekteeritud piltidelt selgelt näha, missugused klassid millises pea piirkonnas domineerivad. Informatsioon kudede jaotuse kohta võib kasuks tulla näiteks modelleerimisel, kus on probleemiks suured andmehulgad. Visualiseerimisel saadud piltide abil on võimalik andmetest eemaldada just need osad, mis pole modelleerimisülesandes kuigi informatiivsed. Antud näite korral on sellisteks osadeks pea tükid, mis koosnevad enamasti vaid hall- ja valgeollusest (vt joonis 12, lk 29 ja joonis 14, lk 31).

KT piltide prognoosimisel MRT piltide põhjal erinevalt käitunud peade erinevused ei tule selgelt esile, kui ühiselt projekteerida nii KT kui ka MRT mõõtmised. Kui aga vaadata MRT mõõtmiste projekteerimisel saadud pilte, tulevad peade erinevused selgemini välja. Põhierinevuseks on MRT andmete projekteeritud punktide hajuvusdiagrammidel antud peade korral see, et Pea 2 korral on võimalik kudede rühmi eristada, ent Pea 1 korral on kudede rühmad omavahel segunenud.

Kasutatud kirjandus

Bishop, C. M. (2009). *Pattern Recognition and Machine Learning.* New York: Springer.

Hinton, G., Roweis, S. (2002). Stochastic neighbor embedding. *Advances in Neural Information Processing Systems*, **15**.

Härdle, W. K., Simar, L. (2015). *Applied Multivariate Statistical Analysis.* Berlin, Heidelberg: Springer.

Khinchin, A. I. (1957). *Mathematical Foundations of Information Theory.* New York: Dover.

Krijthe, J., van der Maaten, L. (2016). *t-distributed stochastic neighbor embedding using a Barnes-Hut implementation.* https://cran.r-project.org/web/packages/Rtsne/Rtsne.pdf (vaadatud 08.04.2017)

Kuljus, K., Bayisa, F., Bolin, D., Lember, J., Yu, J. (2017). *Comparison of hidden Markov chain models and hidden Markov random field models in estimation of computed tomography images.* Arxiv: 1705.01727.

LeCun, Y., Cortes, C., Burges, C. J. C. (1999). *The MNIST database of handwritten digits.* http://yann.lecun.com/exdb/mnist/index.html (vaadatud 01.02.2017)

Lember, J. (2013). Informatsiooniteooria. Loengukonspekt ja ülesanded. Tartu Ülikool. https://www.math.ut.ee/sites/default/files/ms/informatsiooniteooria_kevad_2013.pdf (vaadatud 18.04.2017)

Naidich, T. P., Castillo, M., Cha, S., Smirniotopoulos, J. G. (2012). *Imaging of the Brain: Expert Radiology Series, 1st Edition.* China: Saunders.

van der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, **15**.

van der Maaten, L., Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**.

Lisad

Lisa 1. KT väärtused tavalisemate kudede korral

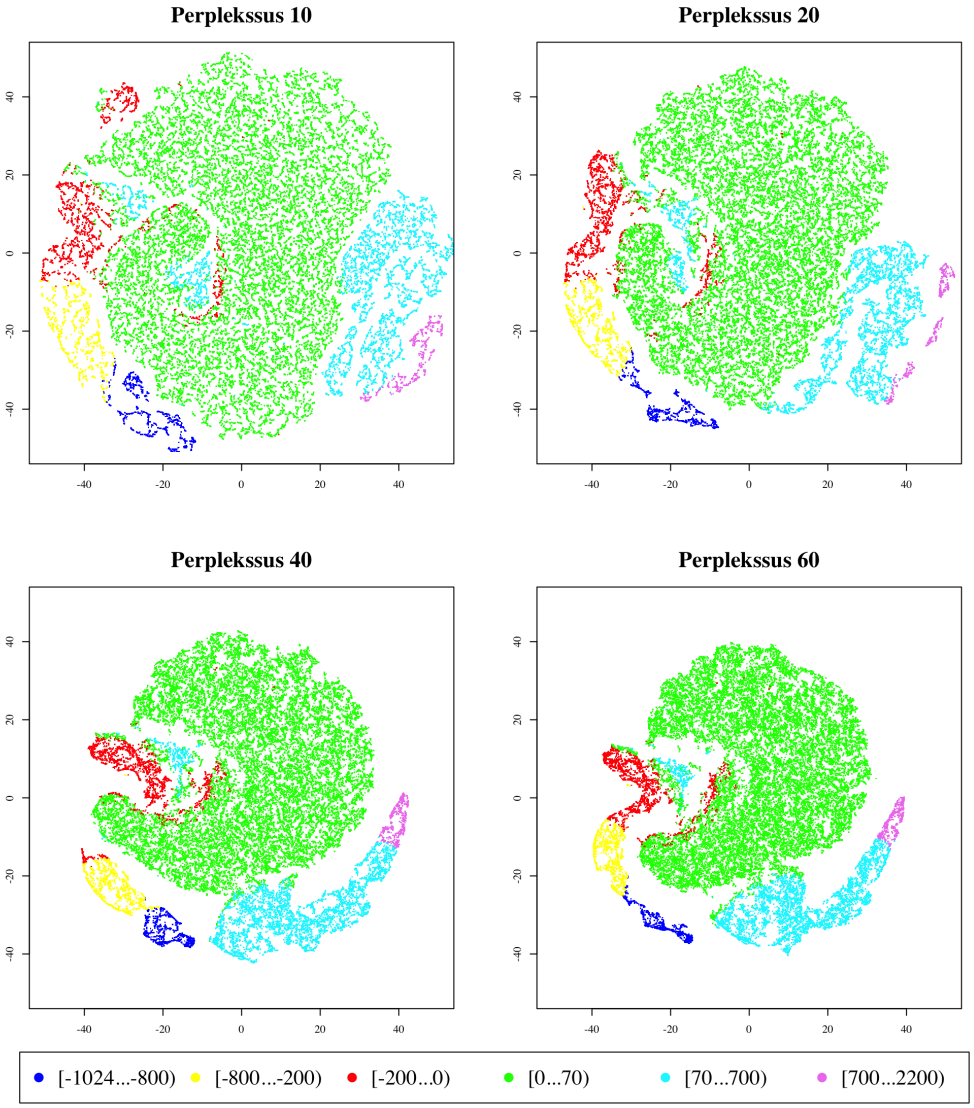
Tabel 4. KT väärtused tavalisemate kudede korral (Naidich et al., 2012, lk 46)

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

Lisa 2. R-koodi näide

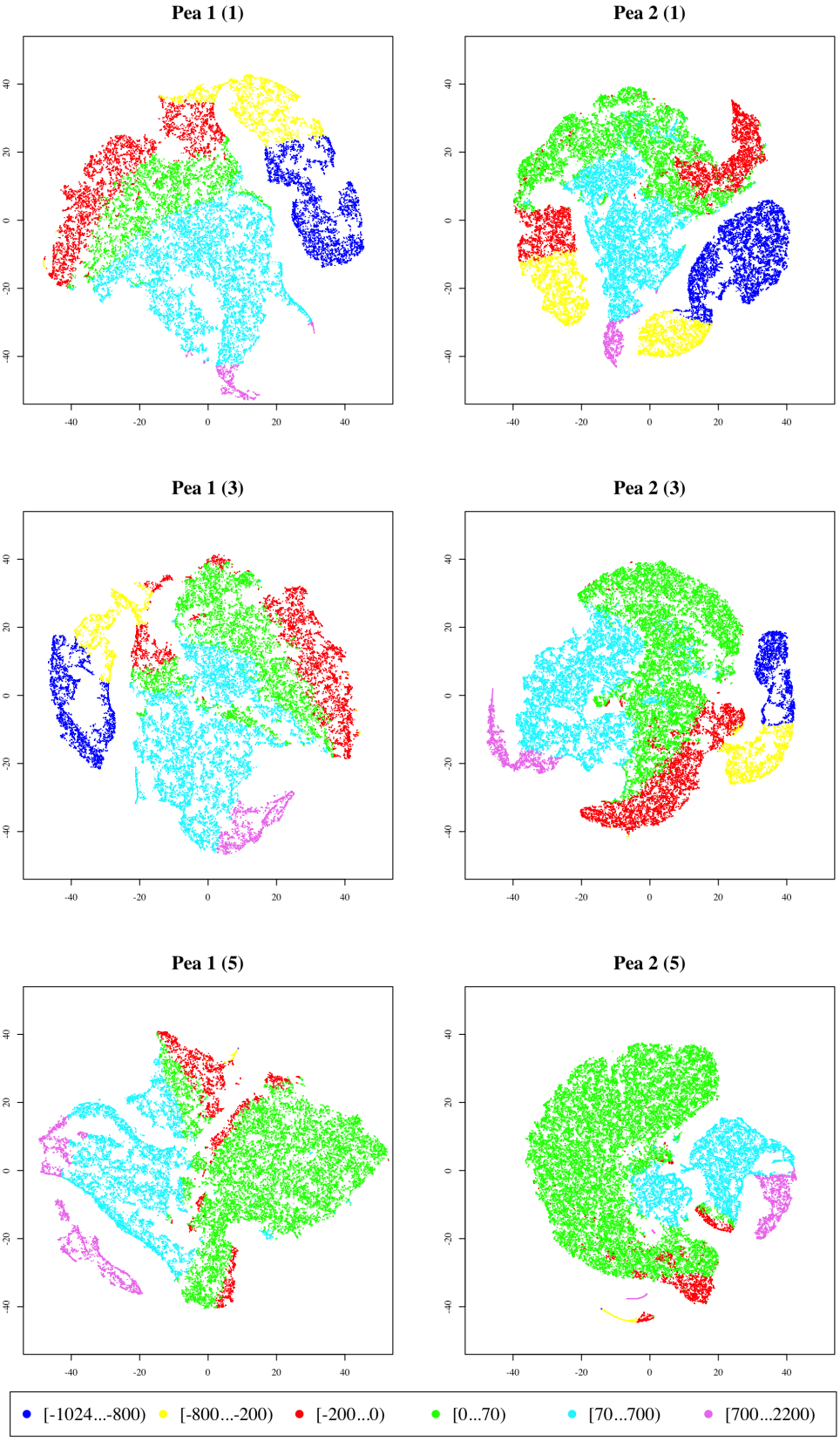
```
library(Rtsne)
set.seed(2)
#Vokslid asuvad pea1_2 ridades. Veerud 2–6 sisaldavad KT ja MRT
#andmeid. Veerus 10 on iga voksli klass.

p10 <- Rtsne(pea1_2[,2:6], pca=FALSE, perplexity=10)

colors <- rainbow(6)
names(colors) <-c("[-1024...-800)","[-800...-200)","[-200...0)",
"[0...70)","[70...700)", "[700...2200)")

plot(p10$Y, t="p",main="Perplekssus 10",xlab="",ylab="",pch=".",
col=colors[pea1_2[,10]])
legend(-55,-90,legend=c(names(colors)),pch=16,col=colors,
horiz=TRUE, cex=0.4,xpd=NA)
```

Lisa 3. Perplekssuse võrdlus pea andmete korral

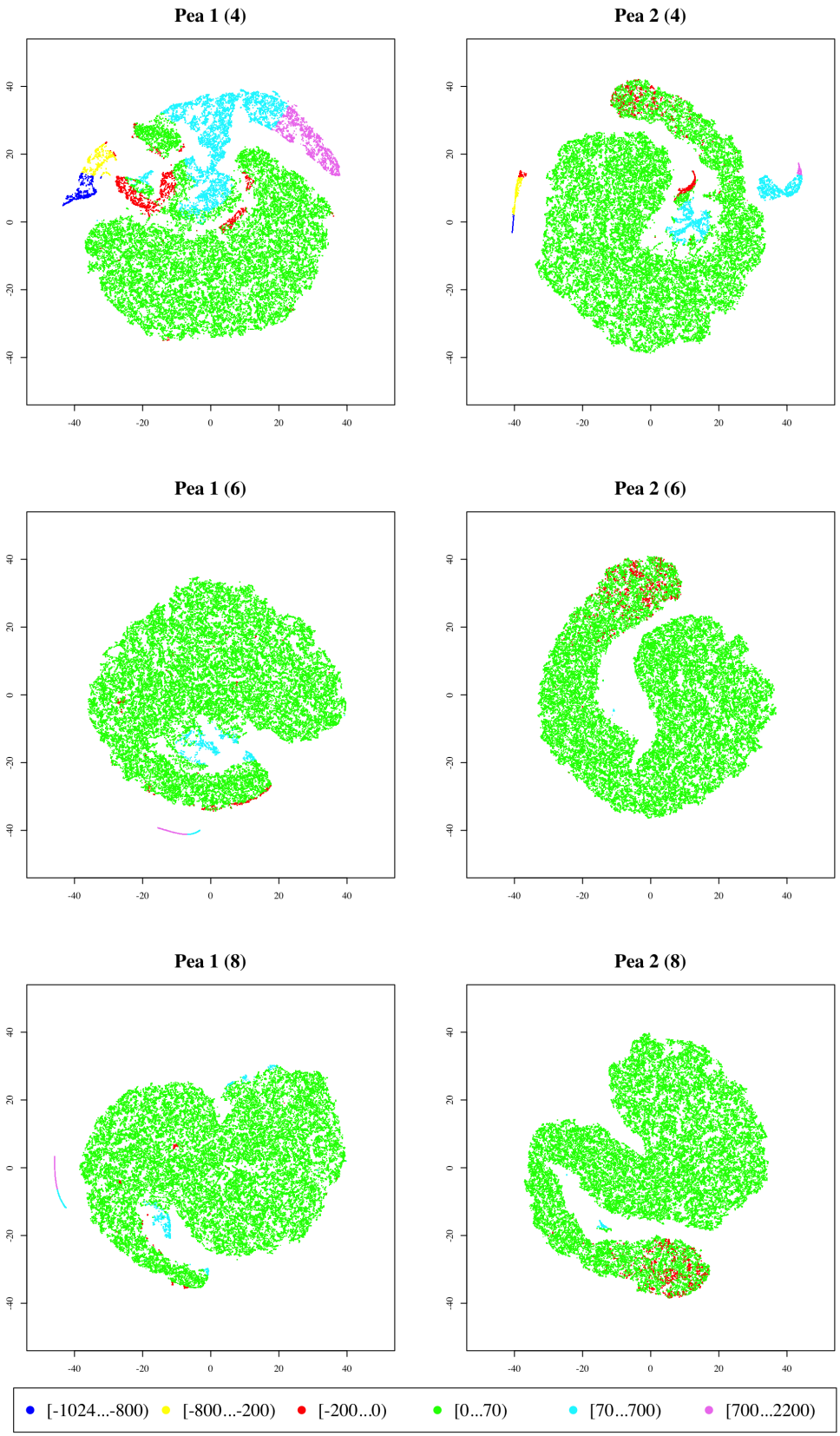


Joonis 10. Pea 1 tüki 2 projekteeritud punktide hajuvusdiagrammid erinevate perplekssuse väärtuste korral

Lisa 4. Pea andmete visualiseerimise tulemused KT ja MRT mõõtmiste korral

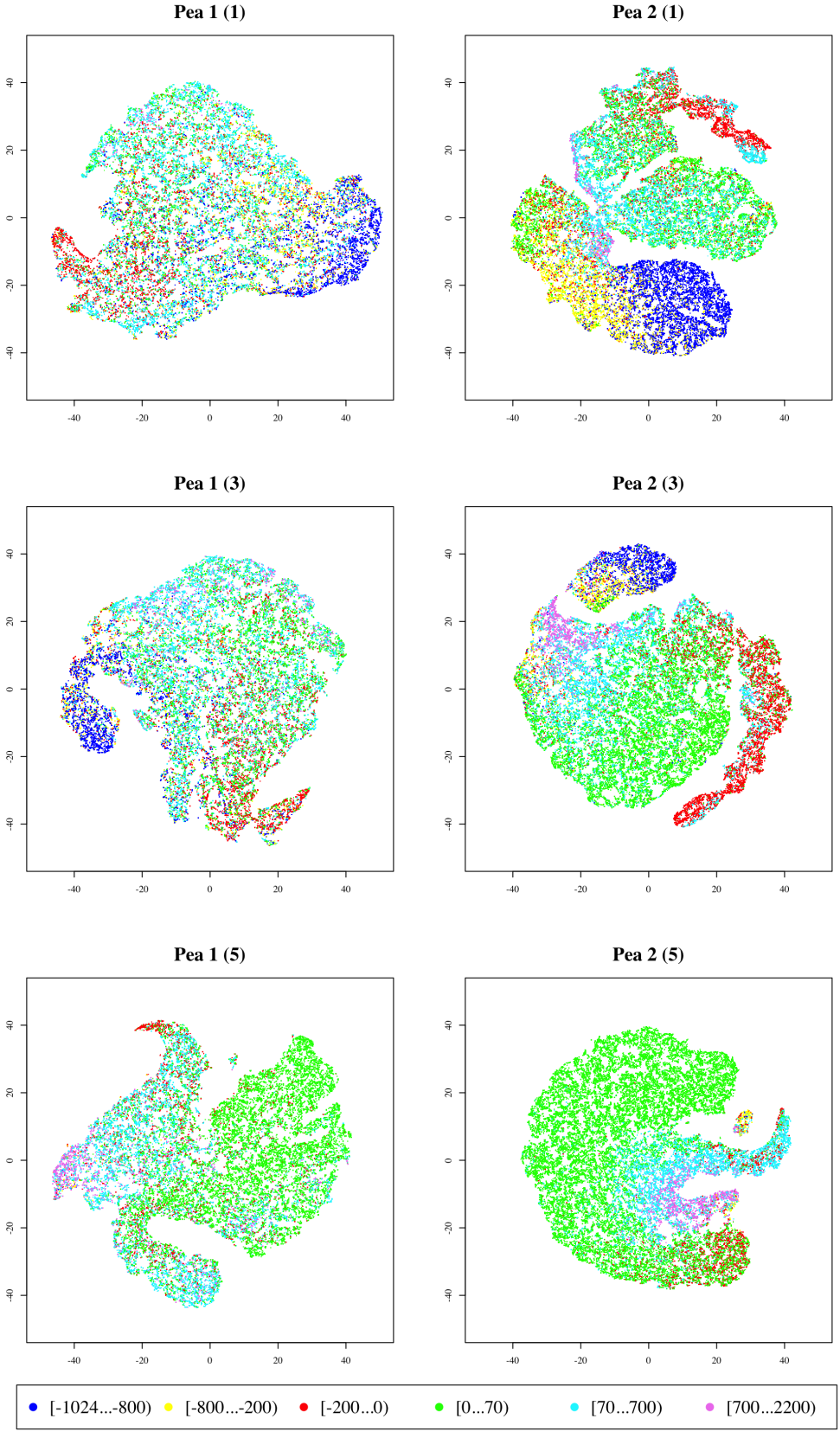


Joonis 11. Pea 1 ja Pea 2 tükkide 1, 3 ja 5 projekteeritud punktide hajuvusdiagrammid KT ja MRT mõõtmiste korral

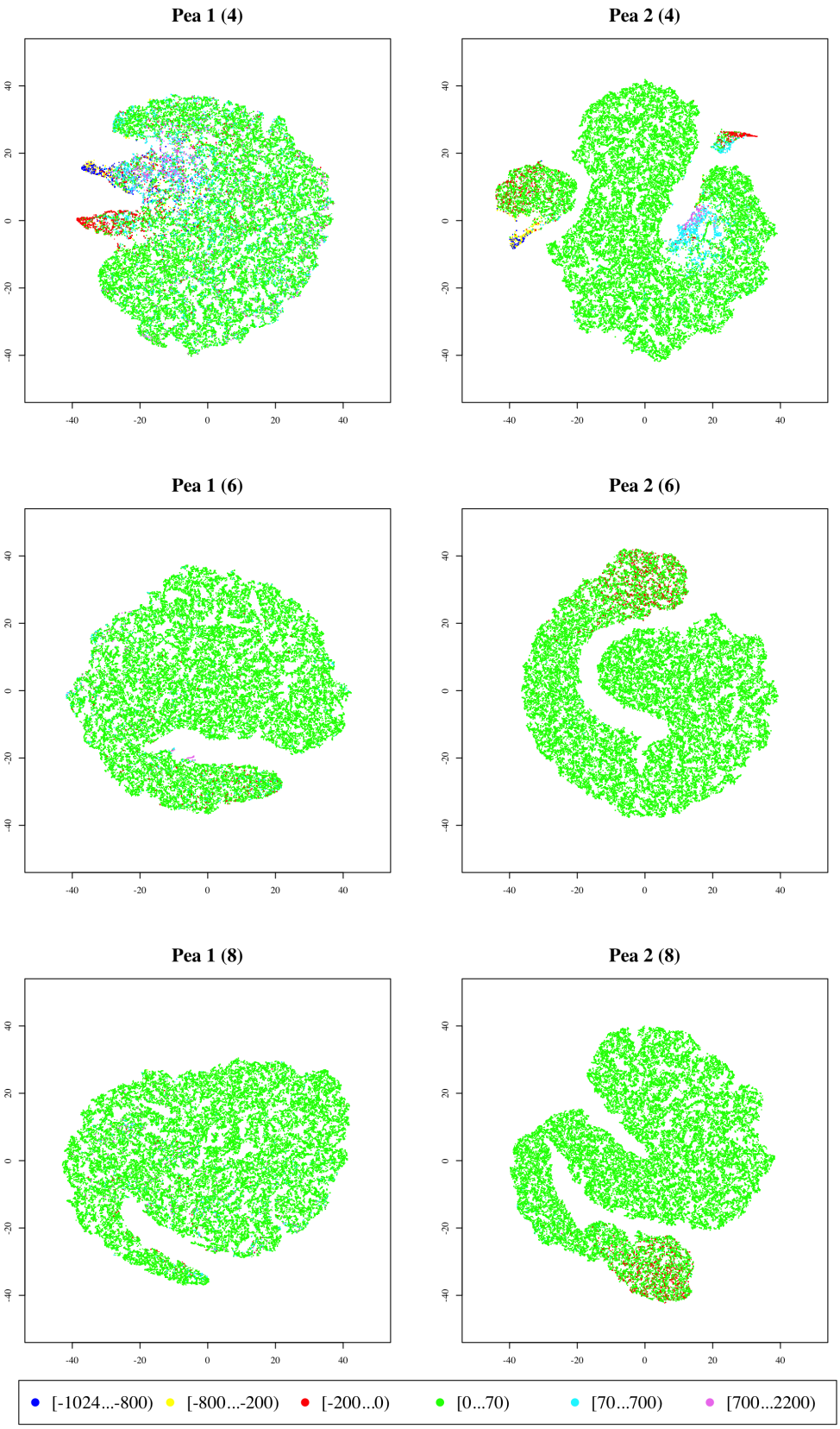


Joonis 12. Pea 1 ja Pea 2 tükkide 4, 6 ja 8 projekteeritud punktide hajuvusdiagrammid KT ja MRT mõõtmiste korral

Lisa 5. Pea andmete visualiseerimise tulemused MRT mõõtmiste korral



Joonis 13. Pea 1 ja Pea 2 tükkide 1, 3 ja 5 projekteeritud punktide hajuvusdiagrammid MRT mõõtmiste korral



Joonis 14. Pea 1 ja Pea 2 tükkide 4, 6 ja 8 projekteeritud punktide hajuvusdiagrammid MRT mõõtmiste korral

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Pirge Kaasik,

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

Mitmemõõtmeliste andmete visualiseerimine meetodi t-SNE abil,

mille juhendaja on Kristi Kuljus,

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

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

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

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

Tartus, **09.05.2017**