Tartu Ülikool
Matemaatika-informaatikateaduskond
Matemaatilise statistika instituut

Kätlin Protsin

Mediaan mitmemõõtmelistes ruumides

Bakalaureusetöö (9 EAP)

Juhendaja: prof. Kalev Pärna

Tartu 2015

Mediaan mitmemõõtmelistes ruumides

Bakalaureusetöö eesmärk on tutvustada mediaani leidmise meetodeid mitmemõõtmelistes ruumides. Erinevalt mediaanist ühemõõtmelises ruumis, on mediaani leidmiseks kõrgemates dimensioonides mitmeid meetodeid ning selles töös on kirjeldatud neist kuute – ruumilise mediaani, marginaalmediaani, Oja mediaani, poolruumi ehk Tukey mediaani, simpleksse sügavuse ehk Liu mediaani ning kumera katte eemaldamise meetodeid. Lisaks on ruumilise mediaani juures toodud välja mõned olulisemad omadused ja nende tõestused.

Märksõnad: keskväärtus, mitmemõõtmeline statistika, robustne statistika

Median in Multidimensional Spaces

The purpose of this Bachelor thesis is to give an overview of methods for finding median in multidimensional spaces. Unlike median in one dimension, there are many methods for finding median in multidimensional spaces and this thesis describes six of them– the spatial median, vector of marginal medians, Oja median, halfspace/Tukey median, simplicial depth/Liu median and convex hull peeling methods. In addition for spatial median there is brought out a few basic properties and their proofs.

Keywords: median, multivariate statistics, robust statistics

Sisukord

Sissejuhatus

Mediaan on statistilises andmetööstuses oluline karakteristik valimi ja üldkogumi kirjeldamiseks. Mediaani üheks parimaks omaduseks on robustsus ehk tema väärtus ei sõltu ekstremaalsetest vaatlustest. Ühemõõtmelise andmestiku mediaani leidmiseks kasutatakse teadatuntud tõsiasja: kui variatsioonireas on paaritu arv liikmeid, on mediaaniks selle rea keskmine liige ning paarisarvu liikmete korral, on mediaaniks kahe keskmise liikme poolsumma.

Bakalaureusetöö eesmärk on anda ülevaade mediaani leidmise meetoditest mitmemõõtmelistes ruumides. Erinevalt mediaanist ühemõõtmelises ruumis, on mitmemõõtmelise mediaani leidmiseks mitmeid erinevaid meetodeid ning antud töös on kirjeldatud neist kuute. Enamus mediaane on defineeritavad optimeerimisülesannete lahenditena, kus on vaja minimeerida sihifunktsiooni või leida maksimaalne sügavus. Mitmemõõtmelisel mediaanil on palju rakendusi tehnikas (2D/3D graafika), majanduses (optimaalse asukoha leidmiseks) ja muudel aladel.

Bakalaureusetöö esimeses peatükis antakse ülevaade mediaanist ühemõõtmelises ruumis. Teises peatükis kirjeldatakse ruumilist mediaani, mis on üks kasutatavamaid meetodeid mediaani leidmiseks kõrgemates dimensioonides. Lisaks on toodud välja ka marginaalmediaani mõiste ja mõned tähtsamad ruumilise mediaani omadused ning nende tõestused. Järgmises peatükis on kirjeldatud alternatiivseid mitmemõõtmelise mediaani määratlusi nagu Oja mediaan, poolruumi ehk Tukey mediaan, simpleksse sügavuse ehk Liu mediaan ja kumera katte eemaldamise meetodil leitud mediaan.

Töös esitatud joonised on tehtud autori poolt ning nende tegemiseks on kasutatud statistikapaketti R.

Autor tänab professor Kalev Pärnat rohkete näpunäidete ja paranduste eest.

1. Mediaan ühemõõtmelises ruumis

Valimi ja üldkogumi olulisteks karakteristikuteks on nn. tsentraalse tendentsi näitajad, millest enim kasutatavad on kaks: keskväärtus (aritmeetiline keskmine) ja mediaan. Mõlemad nimetatud karakteristikud on defineeritavad teatava optimeerimisülesande lahendina.

Esmalt näidatakse, et valimi $X_1, \dots, X_n$ aritmeetiline keskmine on järgmise ruutkaofunktsiooni minimeerimise tulemus

$$\phi(a) = \sum_{i=1}^{n} (X_i - a)^2 \to \min_a.$$

Tõepoolest, teisendades sihifunktsiooni uuele kujule

$$\phi(a) = \sum_{i=1}^{n} (X_i - \bar{X} + \bar{X} - a)^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (X_i - \bar{X})(\bar{X} - a) + n \cdot (\bar{X} - a)^2.$$

Näitame, et teine liidetav on null:

$$\sum_{i=1}^{n} (X_i - \bar{X})(\bar{X} - a) = (\bar{X} - a) \sum_{i=1}^{n} (X_i - \bar{X}) = 0.$$

Samas, esimene avaldis ei sõltu a väärtusest ning seetõttu sihifunktsiooni miinimumi saavutame valikuga $a = \bar{X}$, mis tagab viimase avaldise väärtuse 0.

Definitsioon 1.1 Valimi $X_1, \dots, X_n$ *mediaaniks* nimetatakse arvu m, millest mõlemale poole jääb võrdne arv punkte.

Näide 1. Kui variatsioonireas on paaritu arv liikmeid, on mediaaniks selle rea keskmine liige ($X = 1, 5, 14, 23, 50$, siis $m = 14$). Kui variatsioonireas on paarisarv liikmeid, on mediaaniks kahe keskmise liikme poolsumma ($X = 1, 5, 14, 23, 24, 50$, siis $m = \frac{14+23}{2} = 18{,}5$).

Definitsioon 1.2 Jaotuse mediaan on defineeritud kui

$$P(X \leq m) \geq \frac{1}{2}\,, P(X \geq m) \geq \frac{1}{2}$$

ning jaotuse tihedusfunktsiooni graafikul on mõlemale mediaani poole jäävad pindalad võrdsed.

Selleks, et leida valimi $X_1, \dots, X_n \in \mathbb{R}$ mediaan, tuleb aga minimeerida sihifunktsiooni

$$\phi(m) = \sum_{i=1}^{n} |X_i - m| \to \min_m,$$

mille võib üldisemalt välja kirjutada kui

$$\phi(m) = \int (|X - m|)dx = E(|X - m|) \to \min_m.$$

Kui aga punktid on võetud jaotusest, millel ei eksisteeri lõpliku keskväärtust ehk

$$\int |X|dx = \infty,$$

siis ka

$$\int |X - m|dx = \infty$$

ning sihifunktsiooni minimeerida ei ole võimalik. Selle vältimiseks kasutatakse mugavamat tingimust

$$\int (|X - m| - |X|)dx < \infty.$$

Mediaani üheks parimaks omaduseks on tema robustsus, mis tähendab, et mediaani väärtust ei mõjuta ekstremaalsed punktid. Karakteristiku robustsuse näitajaks on murdepunkt (ingl *breakdown point*).

Definitsioon 1.3 Murdepunktiks nimetatakse kõige väiksemat protsenti muudetud vaatlustest, mis muudaksid funktsiooni hinnangu mistahes suureks. (Tiit, Kollo, Niemi, 1995)

Karakteristik, mille murdepunkt on 0% juures, näiteks aritmeetiline keskmine, on tundlik ekstremaalsete väärtuste suhtes. Ühemõõtmelise mediaani murdepunkt on aga 50% juures, tänu millele võib muuta pooled vaatlustest lõpmata suureks enne kui mediaan muutub. (Tiit, Kollo, Niemi, 1995)

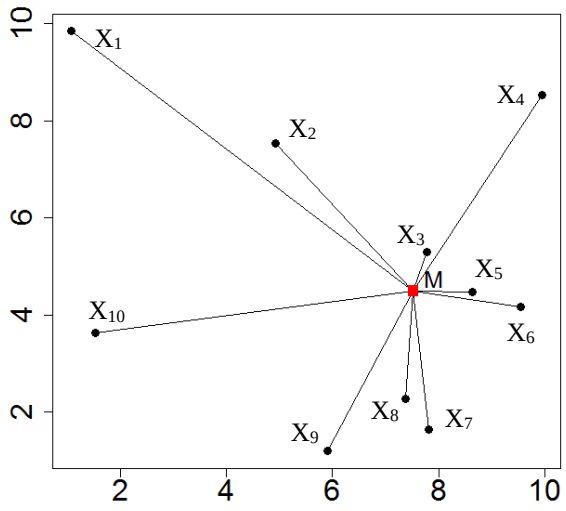
Näide 2. Olgu antud andmekogum $2, 3, 4, 4, 6, 7, 7, 8, 9, 14$. Selle valimi aritmeetiline keskmine on 6,4. Muutes rea ekstreemseima väärtuse 14 mingiks suvaliselt suureks arvuks 1000, on uueks keskmiseks 105. Kuna juba ühe punkti suurendamine muudab aritmeetilise keskmise väärtust, on tema murdepunkt 0% juures.

Sama valimi mediaan on aga 6,7. Muutes samuti rea ekstreemseima väärtuse 14 mingiks suvaliselt suureks arvuks 1000, jääb mediaan samaks. Muutes suuruselt järgmist punkti 9 samuti arvuks 1000, jääb mediaan taas endiseks. Saame nii edasi toimida, kuni pooled valimi väärtustest on muudetud suvaliselt suureks arvuks enne kui mediaan muutub. Seega võime öelda, et mediaani murdepunkt on 50% juures. (Jarman, 2015)

2. Ruumiline mediaan

2.1. Ruumilise mediaani mõiste

Ruumilise mediaani (ka geomeetrilise mediaani või L_1-mediaani) mõiste pärineb Weberi asukoha teooriast. Teooria põhineb oletusel, et üks ettevõte soovib leida optimaalset asukohta laohoone jaoks, mis teenindab n klienti, kelle asukohtadeks on punktid $X_1, X_2, \dots, X_n$. Lahenduse leidmise juures eeldame, et hoone võib ehitada suvalisele koordinaadile ilma piiranguteta. Veel eeldame, et transpordikulud klientideni on võrdelised vastavate kaugustega. Selle asukoha probleemi lahenduseks soovitas Weber minimeerida klientide transpordikulude summa. Kui igale kliendile tehakse hankeid võrdselt, siis laohoone optimaalne asukoht on punkt *M*, mis minimeerib kauguste summa punktist *M* kuni punktideni X_i , $i = 1, 2, \dots n$ (vt. joonis 1). (Small, 1990)



Joonis 1. Ruumiline mediaan kahemõõtmelises ruumis (n=10).

Ülesande üldisemaks püstitamiseks tuuakse sisse vektori normi mõiste.

Definitsioon 2.1.1 Vektori $x \in \mathbb{R}^d$ L_p-normiks ($p > 0$) nimetatakse arvu

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}.$$

Näiteks kahedimensionaalses ruumis, kus $x = (x_1, x_2)$, on vektori L_p-normiks

$$\|x\|_p = (|x_1|^p + |x_2|^p)^{1/p}.$$

Kui $p = 1$, siis normi nimetatakse ka Manhattani normiks:

$$\|x_i\|_1 = |x_{i1}| + |x_{i2}|,$$

kui $p = 2$, siis on normiks Eukleidiline norm

$$\|x_i\|_2 = \sqrt{(x_{i1})^2 + (x_{i2})^2}.$$

(Upton & Cook, 2004)

Nüüd saame defineerida ruumilise mediaani (ingl *spatial median*).

Definitsioon 2.1.2 Punktide $X_1, X_2, \dots, X_n \in \mathbb{R}^d$ *ruumiliseks mediaaniks* nimetatakse vektorit M, mis minimeerib funktsiooni

$$\phi(M) = \sum_{i=1}^{n} \|X_i - M\| \to \min_{M \in \mathbb{R}^d}, \tag{1}$$

kus $\|\cdot\|$ on vektori Eukleidiline norm. (Becker, Fried, Kuhnt, 2013)

Näeme, et $d = 1$ korral taandub definitsioon 1-mõõtmelise mediaani definitsiooniks (iga $p > 0$ korral). Üldistame nüüd definitsiooni 2.1.2 suvalisele jaotusele ruumis $\mathbb{R}^d$.

Definitsioon 2.1.3 Juhusliku vektori $X \in \mathbb{R}^d$ jaotusega $X \sim P$ ruumiliseks mediaaniks nimetatakse vektorit M, mis minimeerib funktsiooni

$$\phi(M) = \int_{\mathbb{R}^d} \|X - M\| P(dx) \to \min_{M \in \boldsymbol{R}^d}.$$

Kahjuks on funktsioon $\phi(M)$ lõplik ainult juhul kui $E\|X\| < \infty$. Sellest kitsendusest vabanemiseks kasutatakse järgmist definitsiooni.

Definitsioon 2.1.4 Juhusliku vektori $X \in \mathbb{R}^d$ jaotusega $X \sim P$ ruumiliseks mediaaniks nimetatakse vektorit M, mis minimeerib funktsiooni

$$\phi(M) = \int_{\mathbb{R}^d} (\|X - M\| - \|X\|) P(dx) \to \min_{M \in \mathbb{R}^d}.$$

Erijuhul, kui tõenäosusmõõt P ruumis $\mathbb{R}^d$ on diskreetne ühtlane jaotus punktidel $X_1, \dots, X_n$, siis definitsioonis 2.1.4 olev kaofunktsioon avaldub kujul

$$\phi(M) = \sum_{i=1}^{n} (\|X_i - M\| - \|X_i\|) \cdot \frac{1}{n} = \frac{1}{n} \left[\sum_{i=1}^{n} \|X_i - M\| - \sum_{i=1}^{n} \|X_i\| \right] \to \min_M.$$

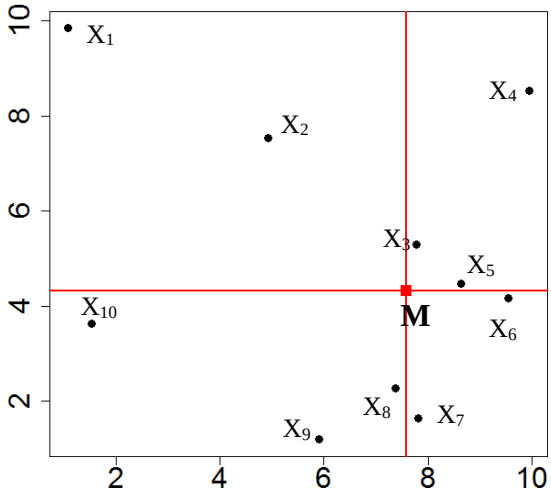
Kuna avaldise viimane summa on konstant (ei sõltu suurusest *M*), siis ülesanne taandub mediaani leidmise ülesandele (1). Sama arutelu kehtib ka suvalise jaotuse korral, kus $\int \|x\| P(x) < \infty$, siis ruumilise mediaani definitsioon 2.1.4 taandub definitsiooniks 2.1.3

$$\phi(M) = \int (\|X - M\| - \|X\|) P(dx) = \int (\|X - M\|) P(dx) - \int (\|X\|) P(dx) \to \min_M.$$

Erijuhul kui $p = 1$, saame nn. *marginaalmediaani*, mille puhul võime minimeerida iga liidetavat eraldi. Näiteks kahemõõtmelises ruumis

$$M(X_1, X_2, \dots, X_n) = \min_{M \in \mathbb{R}^2} \sum_{i=1}^{n} \|X_i - M\| = \min_{m_1} \sum_{i=1}^{n} |x_{i1} - m_1| + \min_{m_2} \sum_{i=1}^{n} |x_{i2} - m_2|.$$

Seega, minimeerides iga liidetavat eraldi, saab marginaalmediaani mitmemõõtmelises ruumis viia ühemõõtmelisse ruumi (vt. joonis 2). (Becker, Fried, Kuhnt, 2013)



Joonis 2. Marginaalmediaan kahemõõtmelises ruumis (n=10).

2.2. Ruumilise mediaani omadused

Olgu $E = \mathbb{R}^d$ Eukleidiline ruum, kus Eukleidiline norm on tähistatud $\|\cdot\|$. Punktide $X_1, X_2, \dots, X_n \in \mathbb{R}^d$ ruumiline mediaan on definitsiooni 2.1.2 kohaselt vektor, mis minimeerib funktsiooni

$$\phi(M) = \sum_{i=1}^{n} \|X_i - M\| \to \min_{M \in \mathbb{R}^d}.$$

Kuna mediaan on punkt, mis minimeerib andmestiku *n* punkti Eukleidiliste kauguste summa, siis selle leidmiseks võib kasutada gradientmeetodit. Funktsiooni gradiendiks nimetatakse selle osatuletiste vektorit ning gradient näitab muutuse kiireima kasvamise suunda. (Clapham & Nicholson, 2005) Järgnevalt tuuakse välja teoreem funktsiooni gradiendi leidmiseks.

Teoreem 2.2.1 Funktsiooni

$$\phi(Y) = \sum_{i=1}^{n} \|X_i - Y\|$$

gradient on

$$\nabla\phi(Y) = -\sum_{i=1}^{n} \frac{(X_i - Y)}{\|X_i - Y\|}.$$

Tõestus. Leitakse funktsiooni $\phi(Y)$ osatuletis kohal Y_j. Kasutades Eukleidilise normi definitsiooni

$$\|X\| = \sqrt{\sum_{k=1}^{d} x_k^2}$$

saadakse

$$\frac{\partial}{\partial Y_j} \sum_{i=1}^{n} \|X_i - Y\| = \frac{\partial}{\partial Y_j} \sum_{i=1}^{n} \left(\sum_{j=1}^{d} \left(X_{ij} - Y_j\right)^2 \right)^{1/2},$$

millest

$$\sum_{i=1}^{n} \left[\frac{1}{2} \left(\sum_{j=1}^{d} \left(X_{ij} - Y_j\right)^2 \right)^{-\frac{1}{2}} \cdot 2\left(X_{ij} - Y_j\right) \cdot (-1) \right] = \sum_{i=1}^{n} \frac{-(X_{ij} - Y_j)}{\left(\sum_{j=1}^{d} \left(X_{ij} - Y_j\right)^2 \right)^{1/2}}$$

$$= \sum_{i=1}^{n} \frac{-(X_{ij} - Y_j)}{\|X_i - Y\|}.$$

Seda lihtsustades saadaksegi funktsiooni $\phi(Y)$ gradient

$$\nabla\phi(Y) = -\sum_{i=1}^{n} \frac{(X_i - Y)}{\|X_i - Y\|}.$$

□

Geomeetrilise mediaani definitsiooni põhjal on teada, et punktide $X_1, X_2, \ldots, X_n$ mediaan M minimeerib funktsiooni ϕ ja arvestades seda on võimalik leida alternatiivne omadus mediaanile. (Bruce, 2011)

Järeldus 2.2.1 Punkt $M \in \mathbb{R}^d$ on valimi $X_1, X_2, \ldots, X_n$ mediaaniks siis ja ainult siis, kui

$$\sum_{i=1}^{n} \frac{(X_i - M)}{\|X_i - M\|} = 0.$$

Tõestus. Minimeerides funktsiooni $\phi(Y)$ ning kasutades teadmist, et funktsiooni gradient lokaalses miinimumis on võrdne nulliga saadakse, et $\nabla\phi(Y) = 0$. Teoreemi 2.2.1 põhjal on $\nabla\phi(Y) = 0$, kui

$$\sum_{i=1}^{n} \frac{(X_i - Y)}{\|X_i - Y\|} = 0.$$

Kui $M \in \mathbb{R}^d$ oleks funktsiooni mediaaniks, siis minimeeriks ta funktsiooni $\phi(M)$ ning kasutades uuesti teoreemi 2.2.1 saab järeldada, et $\nabla\phi(M) = 0$.

Kui võtame suvalise $t \in [0,1]$ ja $x, y \in \mathbb{R}^d$, siis kasutades kolmnurga omadust ja homogeensust saadakse

$$\|tx + (1-t)y\| \leq \|tx\| + \|(1-t)y\| = t\|x\| + (1-t)\|y\|.$$

Sellest järeldub, et Eukleidiline norm on kumer funktsioon. Kui on antud kaks kumerat funktsiooni f ja g, siis distributiivsuse, kolmnurga omaduse ja homogeensuse põhjal

$$\begin{aligned}(f+g)(tx + (1-t)y) &= f(tx + (1-t)y) + g(tx + (1-t)y) \\ &\leq tf(x) + (1-t)f(y) + tg(x) + (1-t)g(y) \\ &= t(f+g)(x) + (1-t)(f+g)(y).\end{aligned}$$

Seega on summa kumeratest funktsioonidest samuti kumer. Sellest järeldub, et M minimeerib funktsiooni $\phi(M)$ ning mediaani definitsiooni põhjal on M punktide $X_1, X_2, \ldots, X_n$ mediaan.

□

See omadus näitab huvitavat fakti – mediaani asukoht ei olene sellest, kui kaugel asuvad andmestiku punktid mediaanist, vaid sellest, kuidas on punktid üldiselt jaotatud. Järgmine järeldus näitab, et andmestiku punkte võib liigutada mööda kiiri, mis lähevad mediaanist nende punktideni, ilma et mediaan muutuks. (Bruce, 2011)

Järeldus 2.2.2 Olgu antud punktid $X_1, X_2, \dots, X_n \in \mathbb{R}^d$ ja olgu M nende andmete mediaan. Kui luua uus andmestik asendades iga punkti $X_i, i = 1, \dots, n$ punktiga $t(X_i - M) + M$, kus $t > 0$, siis mediaan uues andmestikus on samuti M.

Tõestus. Olgu M mediaan, siis järeldusest 2.2.1 teame, et

$$\sum_{i=1}^{n} \frac{(X_i - M)}{\|X_i - M\|} = 0. \tag{2}$$

Nüüd luuakse uus punktihulk, kus iga punkt $X_i, i = 1, \dots, n$ on asendatud punktiga $X_i' \equiv t(X_i - M) + M$, kus $t > 0$. Kasutades järeldust 2.2.1 näidatakse, et M on ka uue andmestiku mediaaniks ehk

$$\sum_{i=1}^{n} \frac{(X_i' - M)}{\|X_i' - M\|} = 0. \tag{3}$$

Saadakse

$$\frac{(X_i' - M)}{\|X_i' - M\|} = \frac{\big((t(X_i - M) + M) - M\big)}{\|(t(X_i - M) + M) - M\|} = \frac{\big(t(X_i - M)\big)}{\|t(X_i - M)\|} = \frac{(X_i - M)}{\|X_i - M\|}.$$

Kuna võrrand (3) on võrdne võrrandiga (2) ning kasutades järeldust 2.2.1 saadakse, et M on ka uue andmestiku mediaaniks. (Bruce, 2011)

□

2.3. Ruumilise mediaani ühesus

Olgu $E = \mathbb{R}^d$ $(d \geq 1)$ Eukleidiline ruum, kus Eukleidiline kaugus on $\|\cdot\|$ ja olgu P tõenäosusmõõt ruumist $\mathbb{R}^d$. Iga vektorit, mis minimeerib funktsiooni $\phi(M) = \int (\|x - M\| - \|x\|)P(dx) = E(\|x - M\| - \|x\|)$ nimetatakse mõõdu P ruumiliseks mediaaniks[1].

[1] Samaväärselt kasutatakse ka terminit mitmemõõtmeline mediaan.

Järgmine teoreem, mis on välja toodud Milasevic ja Ducharme 1987.aasta artiklis annab piisava tingimuse ruumilise mediaani ühesuseks.

Teoreem 2.3.1 Kui tõenäosusmõõt P ei ole Eukleidilises ruumis $E = \mathbb{R}^d$ koondunud ühele sirgele, siis on tal ainult üks ruumiline mediaan.

Tõestus. Oletatakse vastuväiteliselt, et mõõdul P on kaks erinevat mediaani M_1, M_2 nii, et $M_1 \neq M_2$ ja olgu l sirge, mis läbib neid. Iga λ, kus $0 < \lambda < 1$, ja iga $x \in E\backslash l$ korral kehtib lihtne võrdus

$$\begin{aligned}\|x - \lambda M_1 - (1-\lambda)M_2\| - \|x\| &= \|\lambda x + (1-\lambda)x - \lambda M_1 - (1-\lambda)M_2\| - \|x\| \\ &= \|\lambda(x - M_1) + (1-\lambda)(x - M_2)\| - \|x\|.\end{aligned}$$

Kasutades kolmnurga omadust võib väita, et

$$\begin{aligned}&\|\lambda(x - M_1) + (1-\lambda)(x - M_2)\| - \|x\| \\ &\qquad \leq \lambda(\|x - M_1\| - \|x\|) + (1-\lambda)(\|x - M_2\| - \|x\|).\end{aligned} \tag{4}$$

Kui võrratuse (4) mõlemad pooled võrduksid, oleks vektor $\lambda(x - M_1)$ vektori $(1-\lambda)(x - M_2)$ skalaarkordne ning see tähendaks, et $x \in l$, mis oleks vastuolus eeldusega, et $x \in E\backslash l$.

Sellest järeldub, et $M_1 \neq M_2$, $0 < \lambda < 1$ ja iga $x \in E\backslash l$ korral kehtib range võrratus

$$\begin{aligned}&\|\lambda(x - M_1) + (1-\lambda)(x - M_2)\| - \|x\| \\ &\qquad < \lambda(\|x - M_1\| - \|x\|) + (1-\lambda)(\|x - M_2\| - \|x\|).\end{aligned}$$

Kuna tõenäosusmõõt P ei ole koondunud ühele sirgele ehk tõenäosusmõõdu kandja $supp(P) \not\subset l$, siis viimane võrratus kehtib positiivse tõenäosusega $P(\, x \in E\backslash l) > 0$, mistõttu range võrratus jääb kehtima ka siis kui võtta mõlemast poolest keskväärtus[2]

$$\begin{aligned}&E(\|\lambda(x - M_1) + (1-\lambda)(x - M_2)\| - \|x\|) \\ &\qquad < \lambda E(\|x - M_1\| - \|x\|) + (1-\lambda)E(\|x - M_2\| - \|x\|).\end{aligned}$$

[2] Osutub, et kui $Z \leq Y$ ja $P(Z < Y) > 0$, siis $EZ < EY$.
Tõepoolest tõenäosusteooria kursusest on teada, et kui $X \geq 0$ ja $EX = 0$, siis $P(X = 0) = 1$. Sellest järeldub, et kui $X \geq 0$ ja $P(X > 0) > 0$, siis $EX > 0$. Rakendades seda juhul $X = Y - Z$ saadaksegi vajalik võrratus.

Arvestades funktsiooni $\phi(M)$ definitsiooni, on see aga ekvivalentne võrratusega

$$\phi(\lambda M_1 + (1-\lambda)M_2) < \lambda\phi(M_1) + (1-\lambda)\phi(M_2)$$
$$= \lambda \min_M \phi(M) + (1-\lambda) \min_M \phi(M) = \min_{M \in \boldsymbol{R}^d} \phi(M),$$

millest on näha, et punkt $\lambda M_1 + (1-\lambda)M_2$ on sihifunktsiooni $\phi(M)$ mõttes parem punkt võrreldes punktidega M_1 ja M_2. See on aga vastuolus eeldusega, et M_1 ja M_2 on mediaanid.

□

3. Teised mitmemõõtmelise mediaani määratlused

3.1. Oja mediaan

Lisaks ruumilise mediaani meetodile, võib kõrgemates dimensioonides ($d > 1$) mediaani määrata ka Oja meetodil (Oja, 1983). Selleks peaks aga kõigepealt defineerima simpleksi mõiste.

Definitsioon 3.1.1 Simpleksiks nimetatakse $d + 1$ tipulist kumerat keha ruumis $\mathbb{R}^d$. (Clapham & Nicholson, 2005)

Näiteks, ruumis $\mathbb{R}^2$ moodustavad simpleksi kolm punkti, andes tulemuseks kolmnurga. Ruumis $\mathbb{R}^3$ moodustavad simpleksi aga neli punkti ning tulemuseks on tetraeeder. Tippude $T_1, \dots, T_{d+1}$ poolt moodustatud *d*-mõõtmelise simpleksi mahu leidmiseks kasutatakse valemit

$$V(T_1, \dots, T_{d+1}) = \frac{1}{d!}\left|det\begin{pmatrix} 1 & \cdots & 1 \\ t_{11} & \cdots & t_{d+1,1} \\ \vdots & \ddots & \vdots \\ t_{1d} & \cdots & t_{d+1,d} \end{pmatrix}\right|,$$

kus $t_{i1}, \dots, t_{id}$ on punkti T_i koordinaadid, $i = 1, \dots, d + 1$. (Becker, Fried, Kuhnt, 2013)

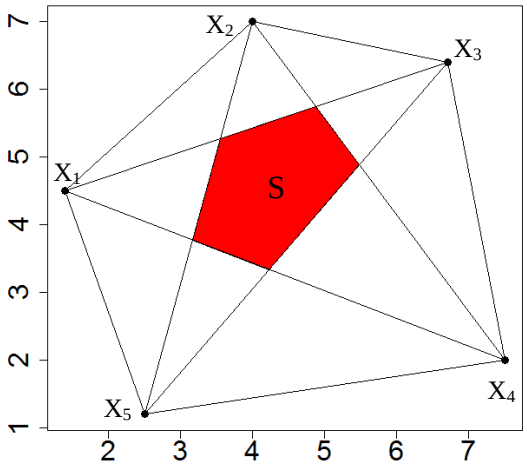
Oja mediaani leidmiseks kasutatavad simpleksid on moodustatud *d* andmepunkti ja oletatava mediaanpunkti *M* vahel.

Definitsioon 3.1.2 Olgu $X = (X_1, \dots, X_n)$ juhuslik valim ruumis $\mathbb{R}^d$. *Oja mediaaniks* nimetatakse vektorit, mis minimeerib sihifunktsiooni

$$\phi(M) = \binom{n}{d}^{-1} \sum_{1 \le i_1 < \cdots < i_d \le n} V\left(X_{i_1}, \dots, X_{i_d}, M\right),$$

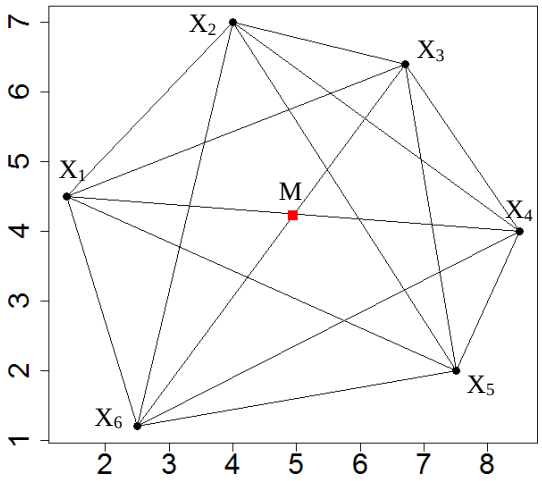
kus $X_{i_1}, \dots, X_{i_d}$ on valimi d punkti ja $\binom{n}{d}^{-1}$ näitab erinevate indeksikomplektide arvu, üle mille summeeritakse. (Becker, Fried, Kuhnt, 2013)

Erinevalt geomeetrilisest mediaanist, ei ole Oja mediaan aga alati ühene. Näiteks, joonisel 3 olevate punktide $X_1, X_2, X_3, X_4, X_5 \in \mathbb{R}^2$ Oja mediaani leidmiseks moodustati kõikvõimalikud simpleksid (kahemõõtmelises ruumis kolmnurgad) kahe andmepunkti X_i ja $X_j, i \neq j$ ja mingi punkti $M \in \mathbb{R}^2$ vahel. Nende simplekside pindalade summa oli minimaalne, kui $M \in S$. Seega, värvitud ala S tähistab kõikvõimalike Oja mediaanpunktide poolt moodustatud kumerat hulka.



Joonis 3. Oja mediaan (värvitud ala) kahemõõtmelises ruumis (n=5).

Joonisel 4 toodud valimipunktide $X_1, \ldots, X_6 \in \mathbb{R}^2$ Oja mediaani leidmiseks minimeeriti samuti kõikvõimalike kolmnurkade pindalade summa. Kuna aga valimipunkte oli paarisarv, oli seekord Oja mediaaniks ühene punkt. (Tiit, Kollo, Niemi, 1995)



Joonis 4. Oja mediaan kahemõõtmelises ruumis (n=6).

3.2. Poolruumi mediaan

Harold Hotelling oli esimene, kes tutvustas poolruumi mediaani mõistet. Ta kirjeldas mediaani kui punkti, mis minimeerib maksimaalse arvu punkte, mis asuvad punktist M „ühel pool“. Ta kujutas seda ette kui kahe jäätisemüüja asukoha probleemi rannajoonel. Hotelling väitis, et optimaalne asukoht esimesel müüjal oleks saabudes valida koht, mis minimeerib inimeste maksimaalse arvu ühel pool.

Kuigi algne idee pärineb Hotellingult oli Tukey see, keda tunnustati selle teooria eest. Kuna poolruumi meetodi interpretatsioon põhineb Tukey sügavuse funktsioonil, on seda mediaani kutsutud ka Tukey mediaaniks. (Liu, Serfling, Souvaine, 2006)

Definitsioon 3.2.1 Hüpertasandiks nimetatakse d-mõõtmelise ruumi $d-1$ mõõtmelist alamruumi. Näiteks ruumis $\mathbb{R}^2$ on hüpertasandiks sirge, kolmemõõtmelise ruumi kahemõõtmeliseks alamruumiks on aga tasand. (Clapham & Nicholson, 2005)

Definitsioon 3.2.2 Poolruumiks nimetatakse hüpertasandi poolt jaotatud ruumi $\mathbb{R}^d$ alamruumi. (Clapham & Nicholson, 2005)

Nüüd on võimalik defineerida poolruumi mediaani leidmiseks vajaliku Tukey sügavuse mõiste.

Definitsioon 3.2.3 Olgu P tõenäosusmõõt ruumist $\mathbb{R}^d$. Punkti $X \in \mathbb{R}^d$ Tukey ehk poolruumi sügavus on defineeritud kui

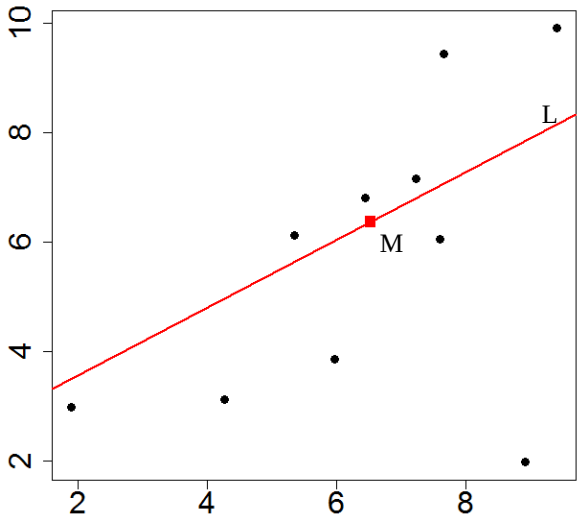
$$D(X|P) = \inf_{X \in H}\{P(H): H \text{ on kinnine poolruum}\}.$$

Tukey sügavus näitab minimaalset arvu andmepunktidest, mis asuvad punkti M läbiva hüpertasandi poolt moodustatud kinnises poolruumis H. (Tukey, 1975)

Definitsioon 3.2.4 Poolruumi ehk *Tukey mediaaniks* nimetatakse punkti $M \in \mathbb{R}^d$, mille korral Tukey sügavus $D(M|P)$ on maksimaalne.

Poolruumi mediaan ei ole tavaliselt ühene punkt. Maksimaalse Tukey sügavusega punktide hulk on kinnine ning tõkestatud kumer hulk. (Becker, Fried, Kuhnt, 2013)

Joonisel 5 olevate vaatluste Tukey mediaan asub punktis M ning seda läbiv kahemõõtmelise ruumi hüpertasand sirge L jaotab ruumi kaheks poolruumiks, kus mõlemasse poolruumi jäävad pooled valimipunktid.



Joonis 5. Hüpertasand (L) ja Tukey mediaan kahemõõtmelises ruumis (n=10).

3.3. Simpleksse sügavuse mediaan

Nagu ütleb ka nimi, põhineb simpleksse sügavuse mediaan (ehk Liu mediaan) simpleksse sügavuse definitsioonil. Tähistagu $x \in \Delta^d$ punkti x kuuluvust d-mõõtmelise ruumi simpleksi ja olgu $I(A)$ sündmuse A indikaatorfunktsioon,

$$I(A) = \begin{cases} 1, kui\ A\ toimub \\ 0, vastasel\ juhul. \end{cases}$$

Kui Oja mediaani leidmise juures kasutatavad simpleksid moodustati d andmepunkti ja mediaanpunkti M poolt, siis Liu mediaani korral moodustavad simpleksi $d+1$ andmepunkti. (Liu, 1990)

Definitsioon 3.3.1 Olgu antud valim $X = (X_1, \dots, X_n)$ ruumis $\mathbb{R}^d$. Punkti x simpleksne sügavus on

$$D(x) = \sum_{1 \le i_1 < \cdots < i_d \le n} I(x \in \Delta^d(X_{i_1}, \dots, X_{i_{d+1}})),$$

kus $\Delta^d(X_{i_1}, \dots, X_{i_{d+1}})$ on $d+1$ andmepunkti poolt moodustatud d-mõõtmeline simpleks.

Simpleksne sügavus näitab punkti x sisaldavate simplekside arvu.

Definitsioon 3.3.2 *Simpleksse sügavuse mediaaniks* nimetatakse punkti M, mille korral simpleksne sügavus $D(M)$ on maksimeeritud.

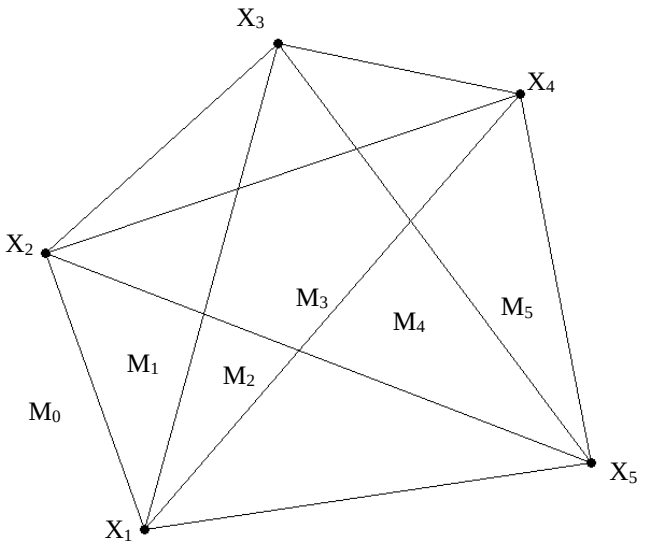
Kui maksimaalse sügavusega punkte on mitu, võetakse mediaaniks nende punktide keskmine. (Liu, 1990)

Simpleksse sügavuse funktsiooni $D(x)$ väärtuste arvutamiseks on kasutusel lihtne meetod:

Olgu L_{ij} sirge, mis läbib valimi X punkte X_i ja X_j , $i \neq j$ ning tähistagu A_{ij} ja B_{ij} poolruume kummalgi pool seda sirget. Kui on võimalik liikuda punktist M_1 punkti M_2 ilma, et peaks ületama sirget L_{ij}, siis $D(M_1) = D(M_2)$. Kui aga punktist M_1 punkti M_2 jõudmiseks on vaja ületada sirget L_{ij}, liikudes nii poolruumist A_{ij} poolruumi B_{ij}, muutub

funktsioon $n_2 - n_1$ võrra, kus n_1 ja n_2 tähistavad valimipunktide arvu vastavalt poolruumis A_{ij} ja B_{ij}. (Tiit, Kollo, Niemi, 1995)

Näide 3.



Joonis 6 Liu simpleksse sügavuse leidmine (n=5).

$D(M_0) = 0$, asub väljaspool valimipunkte;

$D(M_1) = D(M_0) + 3 - 0 = 3$, paremal pool sirget L_{12} (poolruumis B_{12}) on 3 valimi punkti, vasakul (poolruumis A_{12}) 0 punkti;

$D(M_2) = D(M_1) + 2 - 1 = 4$, paremal pool sirget L_{13} (poolruumis B_{13}) on 2 valimi punkti, vasakul (poolruumis A_{13}) 1 punkt;

$D(M_3) = D(M_2) + 2 - 1 = 5$, paremal pool sirget L_{25} (poolruumis B_{25}) on 2 valimi punkti, vasakul (poolruumis A_{25}) 1 punkt;

$D(M_4) = D(M_3) + 1 - 2 = 4$, paremal pool sirget L_{14} (poolruumis B_{14}) on 1 valimi punkt, vasakul (poolruumis A_{14}) 2 punkti;

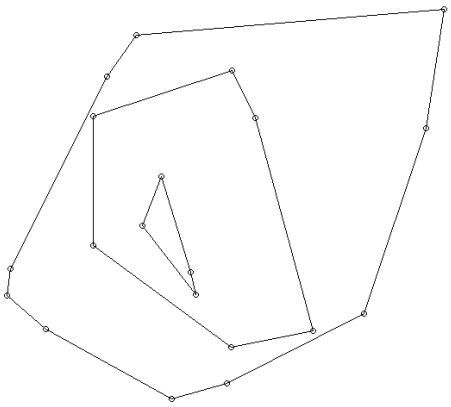
$D(M_5) = D(M_4) + 1 - 2 = 3$, paremal pool sirget L_{35} (poolruumis B_{35}) on 1 valimi punkt, vasakul (poolruumis A_{35}) 2 punkti. (Tiit, Kollo, Niemi, 1995)

3.4. Kumera katte eemaldamise meetod

Kumera katte eemaldamise meetod on visuaalselt kõige parem viis näitamaks mediaani leidmist kõrgemates dimensioonides. Idee põhineb andmestiku keskpunkti leidmisel nii, et eemaldatakse järk-järgult ekstremaalsemad punktid. (Small, 1990)

Definitsioon 3.4.1 Punktihulga $X = (X_1, \dots, X_n)$ kumeraks katteks nimetatakse väikseimat kumerat hulknurka, mis sisaldab kõiki hulga *X* punkte. (Upton & Cook, 2004)

Kumera katte eemaldamise meetodi puhul moodustatakse esmalt punktihulga kumer kate ning seejärel eemaldatakse hulgast X selle katte tipud (eksremaalsemad punktid). Seda protsessi jätkatakse nii kaua kuni allesjäänud punktid on oma kumera katte tippudeks (vt. joonis 7). Kui alles on jäänud ainult üks punkt, ongi see punkt mediaaniks ning kui kumer kate koosneb rohkem kui ühest punktist, on mediaaniks nende punktide raskuskese (aritmeetiline keskmine). (Small, 1990)



Joonis 7. Kumera katte eemaldamise meetod kahemõõtmelises ruumis (n=20).

Üheks enim kasutatavamaks algoritmiks punktide eemaldamiseks kumerast kattest, on Grahami skaneering, mis leiab punktihulga ekstreemsed punktid aja $O(n \log n)$ jooksul. Grahami meetodi puhul leitakse kõigepealt kõige madalama *y*-koordinaadiga punkt *P* (kui neid punkte on mitu, tuleks valida see, millel on suurem *x*-koordinaat). Seejärel järjestatakse ülejäänud punktid nurga (mis tekib punkti ja *P* vahel) suuruse alusel

järjekorda alates väiksemast. Edasi konstrueeritakse kumer kate, kus läbitakse punktid järjekorras ja lisatakse kattele, kui tehakse pööre vasakule (kellaosuti liikumisele vastupidises suunas). (Ramaswami, 1993)

Veel on üheks algoritmiks hulga ekstreemsete punktide eemaldamiseks Jarvise „kinkepakendi“ (ingl *gift-wrapping*) algoritm. Nagu Grahami meetodi puhul, on kõigepealt vaja leida kõige minimaalse *y*-koordinaadiga punkt *P*. Järgmiseks valitakse kattesse punkt, mis moodustab punktiga *P* kõige väiksema polaarnurga ning ülejäänud katte tipud on leitud kõige väiksema polaarnurga alusel eelmisest punktist. Jarvise algoritmi kiirus sõltub katte tippude arvust h ning lahendub tavaliselt aja $O(nh)$ jooksul. (Ramaswami, 1993)

Kokkuvõte

Bakalaureusetöö eesmärk oli anda ülevaade erinevatest mediaani leidmise meetoditest mitmemõõtmelistes ruumides.

Töö esimeses peatükis kirjeldati nii valimi kui jaotuse mediaani leidmist ühemõõtmelises ruumis ja toodi välja mediaani robustsust kirjeldav näitaja murdepunkt.

Teises osas tutvustati põhjalikumalt enamlevinuimat mitmemõõtmelist mediaani – ruumilist mediaani, mis põhineb sihifunktsiooni minimeerimisel. Lühidalt on kirjeldatud ka marginaalmediaani, mille saab mitmemõõtmelisest ruumist viia ühemõõtmelisse ruumi. Lisaks toodi välja ruumilise mediaani leidmise viis gradientmeetodil ning paar põhilist ruumilise mediaani omadust koos tõestustega. Sealjuures tõestati, et ruumiline mediaan on ühene.

Bakalaureusetöö kolmandas peatükis kirjeldati lähemalt alternatiivseid mitmemõõtmelise mediaani määratlusi. Oja mediaani leidmiseks tuleb minimeerida andmepunktide ja oletatava mediaanpunkti poolt moodustatud simplekside mahtude summa. Poolruumi mediaani jaoks on vajalik leida punkt, mille korral Tukey sügavus on maksimeeritud. Simpleksse sügavuse mediaani korral tuleb maksimeerida aga punkti simpleksset sügavust. Viimasena kirjeldatud kumera katte eemaldamise meetodi puhul tuleb mediaani leidmiseks järk-järgult eemaldada punktihulgast ekstremaalsemad punktid.

Töös oli toodud välja ka joonised, kus iga meetodi puhul oli leitud juhusliku valimi põhjal mediaan kahemõõtmelises ruumis.

Kasutatud kirjandus

1) Becker, C., Fried, R., Kuhnt, S. (Eds.). (2013). *Robustness and Complex Data Structures*. Berlin, Heidelberg: Springer Berlin Heidelberg. Kättesaadav: http://link.springer.com/10.1007/978-3-642-35494-6 (26.04.15)
2) Bruce, D. (2011). A Multivariate Median in Banach Spaces and Applications to Robust PCA. Kättesaadav: http://www-personal.umich.edu/~romanv/students/bruce-REU.pdf (26.04.15)
3) Clapham, C. & Nicholson, J. (2005). *The concise Oxford dictionary of mathematics* (3rd ed). Oxford: Oxford University Press.
4) Jarman, K. H. (2015). *Beyond Basic Statistics: Tips, Tricks, and Techniques Every Data Analyst Should Know*. John Wiley & Sons.
5) Liu, R. Y. (1990). On a Notion of Data Depth Based on Random Simplices. *The Annals of Statistics*, *18*(1), 405–414.
6) Liu, R. Y., Serfling, R. J., Souvaine, D. L. (2006). *Data Depth: Robust Multivariate Analysis, Computational Geometry, and Applications*. American Mathematical Soc.
7) Milasevic, P., Ducharme, G. R. (1987). Uniqueness of the Spatial Median. *The Annals of Statistics*, *15*(3), 1332–1333. Kättesaadav: http://doi.org/10.1214/aos/1176350511 (26.04.15)
8) Oja, H. (1983). Descriptive Statistics fo Multivariate Distributions. *Statistics & Probability Letters, 1, 327-332.*
9) Ramaswami, S. (1993). Convex Hulls: Complexity and Applications (A Survey). Kättesaadav: http://repository.upenn.edu/cis_reports/264/ (26.04.15)

10) Small, C. G. (1990). A Survey of Multidimensional Medians. *International Statistical Review / Revue Internationale de Statistique*, *58*(3), 263–277. Kättesaadav: http://doi.org/10.2307/1403809 (26.04.15)

11) Tiit, E.-M., Kollo, T., Niemi, H. (1995). *Multivariate Statistics and Matrices in Statistics: Proceedings of the 5th Tartu Conference, Tartu, Pühajärve, Estonia, 23-25 May 1994*. BRILL.

12) Tukey, J. W. (1975). Mathematics and Picturing Data. In R. James (Ed.), *Proceedings of the 1974 international congress of mathematicians,* Vancouver (Vol. 2, pp. 523-531).

13) Upton, G. J. G. & Cook, I. T. (2004). *A dictionary of statistics*. Oxford: Oxford University Press.

Lisa 1. R-i kood jooniste jaoks

```
####L1 MEDIAAN
##Hetkel leitud kahemõõtmelise ruumi jaoks, kuid võimalik ka kõrgemates
dimensioonides
library(pcaPP)
n=10
x1=c(runif(n,1,10))
x2=c(runif(n,1,10))
#Teeme mediaani jaoks maatriksi
x=cbind(x1,x2)
med=c(l1median(x))
plot(x1,x2,pch=21,bg="black",cex=1.3,cex.axis=2,cex.lab=2)
for (i in (1:n)){
      segments(med[1],med[2],x1[i],x2[i])
}
points(med[1],med[2],pch=15,col="red",cex=1.6,bg="red")

####Marginaalmediaan
##Võimalik leida ka kõrgemates dimensioonides
plot(x1,x2,pch=21,bg="black",cex=1.3,cex.axis=2,cex.lab=2)
#Leiame mediaani x-teljel ja y-teljel
med1=median(x1)
med2=median(x2)
abline(v=med1,col="red",lwd=2)
abline(h=med2,col="red",lwd=2)
#Mediaan
points(med1,med2,pch=15,col="red",cex=1.6,bg="red")

####Oja mediaan
##Töö joonisel on kasutatud library OjaNP demo
##Funktsioon ojaMedian annab vastuseks ühe kõikvõimalikest Oja
mediaanväärtustest
##Võimalik leida ka kõrgemates dimensioonides
library(OjaNP)
n=5
x11=c(runif(n,1,10))
x22=c(runif(n,1,10))
xx=cbind(x11,x22)
oja=c(ojaMedian(xx))
plot(x11,x22,pch=21,bg="black",cex=1.3,cex.axis=2,cex.lab=2)
points(oja[1],oja[2],col="red")

####Tukey mediaan
##Võimalik arvutada ka kõrgemates dimensioonides, aga siis on tulemus
ligikaudne
##Kõikvõimalike väärtuste kõige keskmiseim väärtus
library(depth)
tuk=med(x)
plot(x1,x2,pch=21,bg="black",cex=1.3,cex.axis=2,cex.lab=2)
tuk=unlist(tuk)
```

```
points(tuk[1],tuk[2],pch=15,cex=1.6,col="red")
#Hüpertasand
nmod=lm(I(x2-tuk[2])~I(x1-tuk[1])+0)
abline(predict(nmod, newdata=list(x1=0))+tuk[2], coef(nmod),
col='red',lwd=2)

####Liu mediaan
##Võimalik ainult kahemõõtmelistes ruumides
##Kõikvõimalike väärtuse kõige keskmisem väärtus
liu=med(x,method="Liu")
plot(x1,x2,pch=21,bg="black",cex=1.3,cex.axis=2,cex.lab=2)
liu=unlist(liu)
points(liu[1],liu[2],col="red")

####Sama funktsiooniga med() on võimalik leida ka Oja mediaani (ainult
2D),
####Spatial mediaani (ka kõrgemates dimensioonides)

####Kumera katte eemaldamine
peel=x
plot(peel)
hpts1=chull(peel)
hpts2=c(hpts1,hpts1[1])
lines(peel[hpts2, ])
peel=peel[-hpts1,]
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Kätlin Protsin

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Mediaan mitmemõõtmelistes ruumides“, mille juhendaja on Kalev Pärna
   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.

Tartu, **28.04.2015**