

Mitmemõõtmeline statistika

- Mida me mõtleme mitmemõõtmelise statistika all?
 - mitu **mõõdet** e mitu **tunnust** e mitu **omadust**, mis kirjeldavad meid huvitavaid objekte
- Andmete tekkimise/kirjeldamise taga on mitmemõõtmeline struktuur
 - Palju omadusi, mis aitavad sarnasusi/erinevusi välja tuua

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

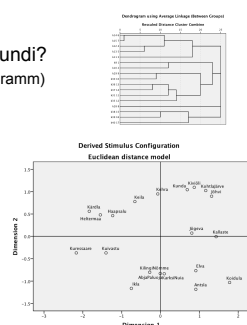
Mitmemõõtmeline skaleerimine

Tunnuste ja objektide GRAAFILINE grupeerimine

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Mitmemõõtmeline skaleerimine

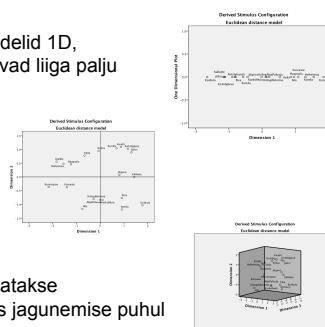
- Milliste grupeerimise meetodite juures saime veel graafilise väljundi?
 - hierarhiline klasteranalüüs (dendrogramm)
- Skaleerimise põhitulem on **graafiline lahendus**
 - mõõtmete taandamine
 - peaks aitama paremini tajuda andmete taga olevat struktuuri



Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Mudelid

- Erandjuhul on mudelid 1D, kaugused moonduvad liiga palju
- Reeglina tehakse 2D mudelid
- 3D mudeleid kasutatakse selgete gruppideks jagunemise puhul

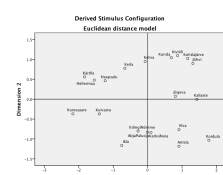


Mitmemõõtmeline statistika Kairi Osula 2013/kevad

2D

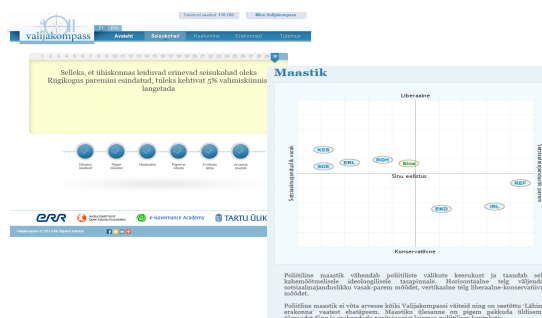
- Parim mudel – **kahemõõtmeline**
- Andmed **jagunevad** kahe uue mõõtmelise/dimensiooni vahel
- Tekivad** grupid/rühmad

- Kuid võib olla ka nii, et ma loodan, et andmed grupeeruvad, kuid kui olen kasutanud mitut meetodit ja ikka **ei teki selget gruppi**.
- Samas ma olen saanud (nt. seitsmest tunnusest) kaks tunnust. Saan kirjeldada oma objektid nende kahe kaudu ja lisaks näen, millised tunnused domineerivad jne.



Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Näide (valijakompass.err.ee)



Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Veel näiteid

- iMapper <http://cl.tlu.ee/INI/>
- iCamp <http://www.htk.tlu.ee/iFolio/>

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Põhimõisted

- Mitmemõõtmeline skaleerimine – MDS**
(ingl. k. *Multi-dimensional Scaling*) on tehnika, mida kasutatakse mitmemõõtmeliste ehk paljude tunnuste abil kirjeldatud või kirjeldatavate andmete struktuuri võimalikult täpselt ja lihtsaks graafiliseks esitamiseks (reeglina ühe-, kahe- või kolmemõõtmelisel kujul).
- Kaugus ehk erinevus** – (ingl.k. *dissimilarity*)
- Lähedus ehk sarnasus** - (ingl.k. *similarity*)
- Koormus** - (ingl.k. *stress*) on mõõt, mis kirjeldab lähteandmete (originaalkauguste) ja MDS'i kaugus saadud lahenduse (tuletatud kauguste) omavahelist erinevust. Mida väiksem koormus, seda parem vastavus on algandmete ja lahendi vahel.

K.Niglas (loenguslaidid)

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Koormus

- Kui hästi meie väiksema arvu mõõtmega lahendus/ mudel taastoodab sarnaseid kauguseid?
- Millist mudelit meie tahaks?
 - Sellist, mille koormus on võimalikult suur
 - Sellist, mille koormus on võimalikult väike

Koormuse näitajate tõlgendamine

Goodness of fit (Kruskal)

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Koormus

- Kuidas on seotud **mõõtmed** ja **mudeli headus**?
 - Mida vähem mõõtmel me alles jätame, seda rohkem me surume kokku esialgseid tegelikke erinevusi/sarnasusi ja seda suurem koormus sellele mudelile tuleb.
- Teen mitu mudelit (1D, 2D, 3D jne) ja hakkan võrdlema **koormuse taseme muutut**.
 - Kui lisan ühe mõõtme juurde, vaatan, kui palju koormus muutub
 - Iga mõõtme lisandumisega läheb koormus **väiksemaks**
 - Kas see efekt, millega ma lisan mudelisse ühe mõõtme juurde, kas see tasub mudeli headuses ennast ära?
 - Mingil sammul näen, et loodud mudel on juba liiga moonutatud (ja seda ei saa kasulikuks pidada)
 - Mingi osa alginfost läheb kaduma. Poleks ju vaja olnud mõelda andmete taha viite mõõdet, kui need on kahe mõõtmega mõõdetavad.

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Mudeli headuse tõlgendamine

- Kõige olulisem – mudeli **interpreteeritavus**
Kas me suudame saadud uutele mõõtmetele anda sisulise tähenduse – niisuguse, mis seletab nähtust ja andmetes olevat struktuuri lihtsamini ja selgemini.
- Ja sinna juurde vaatame **stressi e koormuse väärtust**. Kui koormuse näitaja on hästi suur, siis isegi kui tundub, et mudelit on hästi lihtne interpreteerida ei ole seda mõtet teha, sest see mudel on liiga moonutatud.

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Andmestikud MDS meetodi rakendamisel

- Ruutmaatriksid** – kaugus- või sarnasusmõõtdena esitatud andmed: objektiivsed mõõtmised või subjektiivsed hinnangud konkreetse mõõtmistulemuse väärtus, võrdluste keskmine tulemus, erinevate/sarnaste elementide absoluutarvu või suhtarvu, mis näitab, mitu inimest liigitasid objektid samasse kategooriasse/paari/vms, jne
- Keerulisema struktuuriga maatriksid** – erinevad katserühmad, erinevad alamvalimid, täiendavad andmed kaugus- või sarnasusmõõtdetele lisaks, ...
- Tavapärane mitmemõõtmeline andmestik** - kaugus- või sarnasusmõõtded arvutatakse olemasolevate tunnuste põhjal (nt Eukleidiise kauguse valemist kasutades)

K.Niglas (loenguslaidid)

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLICOOL

Näide 1. linnad.sav

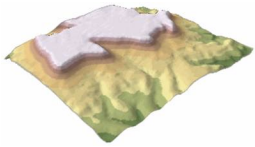
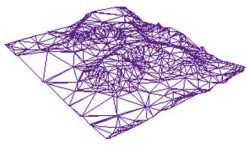
	1	2	3	4	5	6	7	8	9	10
	Abja-Paluoja	Antsla	Elva	Haapsalu	Heltermaa	Ikla	Jõgeva	Jõhvi	Kallaste	Karksi-Nuia
1 Abja-Paluoja		92	85	178	206	105	121	223	160	12
2 Antsla	92		68	270	298	197	137	213	131	80
3 Elva	85	68		262	290	189	90	167	84	72
4 Haapsalu	178	270	262		31	175	219	265	280	190
5 Heltermaa	206	298	290	31		203	250	296	311	218
6 Ikla	105	197	189	175	203		216	302	256	117
7 Jõgeva	121	137	90	219	250	216		104	62	119
8 Jõhvi	223	213	167	265	296	302	104		96	221
9 Kallaste	160	131	84	280	311	256	62	96		158
10 Karksi-Nuia	12	80	72	190	218	117	119	221	158	

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLICOOL

Kaugus linnade vahel

- 2D linnulennult (eukleidiline kaugus)
- Kaugus, mida reaalselt läbime ühest linnast teise, 3D

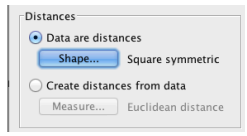




Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLICOOL

Näide 1.

- linnad.sav
- SPSS: Analyze/Scale/Multidimensional Scaling(ALSCAL)
 - Variables (kõik linnad)
 - Data are distances
 - Shape (andmestik on sümmeetriline, asümmeetriline, riskülikukujuline)
 - Create distances from data (SPSS arvutab kaugused)

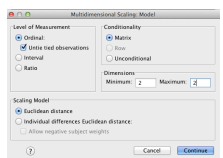


Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLICOOL

Multidimensional Scaling - Model

- Level of measurement
 - Ordinal – kui hakatakse taastootma kauguseid, siis püütakse säilitada kauguste proportsioonid, mitte täpsed kaugused
 - Untie tied observations – kui kahe objekti kaugus on sarnane, siis nende põhjal taastoodetud kaugusmõõt ei ole sarnane
- Scaling Model
 - Euclidean distance
- Conditionality
 - Matrix
- Dimensions
 - Min: ____ Max: ____

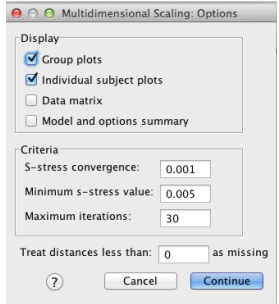


Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLICOOL

Multidimensional Scaling - Options

- Groups plots
- Individual subject plots

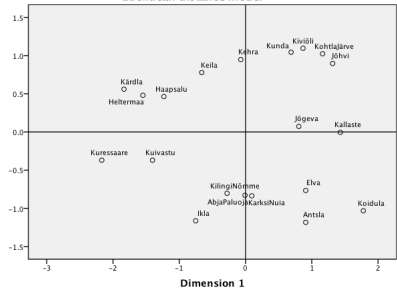


Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLICOOL

Näide 1.

Derived Stimulus Configuration
Euclidean distance model

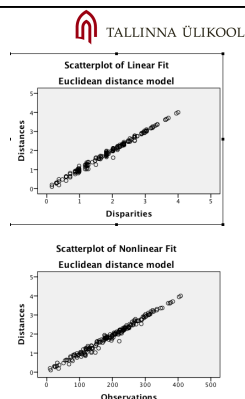


Stimulus Number	Stimulus Name	Dimension 1	Dimension 2
1	Abja-Paluoja	-.0079	-.8297
2	Antsla	-.0094	-1.1863
3	Elva	-.0094	-.7667
4	Haapsalu	-1.1317	-.4629
5	Heltermaa	-1.5491	-.4804
6	Ikla	-1.7522	-1.1639
7	Jõgeva	-.8052	-.0723
8	Jõhvi	1.1147	-.0873
9	Kallaste	1.4338	-.0559
10	Karksi-Nuia	-.5945	-.4368
11	Kelja	-.0709	-.9485
12	Kunda	-.4609	-.7772
13	Kõivaste	-.2793	-.8034
14	Kõivaste	-.4862	1.0959
15	Kõivaste	1.1436	1.0488
16	Kõivaste	1.7783	-1.0323
17	Kõivaste	-1.1490	-.3697
18	Kõivaste	-.4875	1.0439
19	Kõivaste	-1.1490	-.3705
20	Kõivaste	-1.1435	-.5608

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Näide 1.

- Mudeli "headuse" näitajad
- Mida rohkem on punktid ühel joonel koos, seda väiksemad on taastootmise vead.
- Mida rohkem on punktipilv hajuv, seda rohkem on juhuslikkust.



Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Näide 1.

- Koormuse e stressi näitaja

Iteration history for the 2 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

Iteration	S-stress	Improvement
1	.06998	
2	.04992	.02006
3	.04638	.00354
4	.04545	.00093

Kui suur oli koormus ümberarvutamise tsükli ajal. Neljanda pööramise koormus oli 0,04

Iterations stopped because S-stress improvement is less than .001000

Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances. Stress values are Kruskal's stress formula 1.

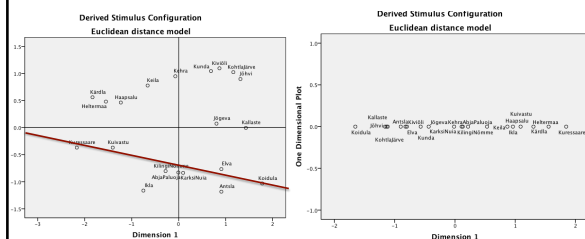
For matrix Stress = .03616 RSQ = .99240 Lõplik stressi näitaja on 0,03

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Näide 2.

2D mudel

1D mudel

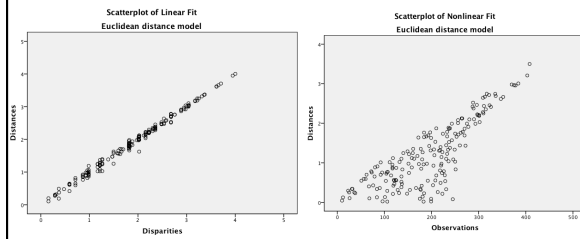


Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Näide 2.

2D mudel

1D mudel



Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Näide 2.

2D mudel

Stressi näitaja 0,003

1D mudel

S = 0,25

For matrix Stress = .25491

RSQ (korrelatsiooni ruut) taastoodetud variatsiooni osa
Enne oli 99% taastoodetud, nüüd on vaid 79%.

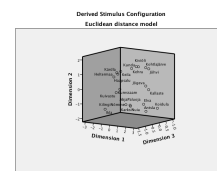
RSQ = .79609

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Näide 3.

- Kas 3D mudel on parem/halvem?

- Pilt selgem/segasem?

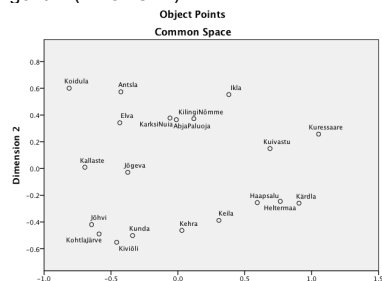


- Koormuse näitaja suurem/väiksem?
S=0,025

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Näide 4.

- Kasutame teist algoritmi (PROXCAL)



Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Kinship.sav

- 15 kaarti
- ~90 objekti moodustasid kaartidest paare, millised kaardid sobivad kokku
- Tabelisse sisestati, kui paljud ei pannud neid kokku paari

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Näide 5

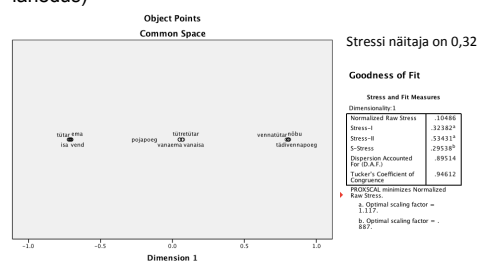
- Mis on need peidetud nähtused, mille kaudu inimesed tajuvad selliste kategooriate tekkimist?

	valik	ladi	vend	nõbu	tütar	isa	tütretütar	vanaema
1	ladi	0	0	0	0	0	0	0
2	vend	83.00	0	0	0	0	0	0
3	nõbu	38.00	77.00	0	0	0	0	0
4	tütar	78.00	61.00	83.00	0	0	0	0
5	isa	79.00	55.00	84.00	43.00	0	0	0
6	tütretütar	79.00	82.00	82.00	69.00	83.00	0	0
7	vanaema	84.00	76.00	84.00	83.00	73.00	48.00	0
8	vanaema	78.00	83.00	84.00	74.00	82.00	38.00	11.00
9	pojapoeg	85.00	73.00	81.00	80.00	74.00	13.00	38.00
10	ema	72.00	63.00	84.00	34.00	13.00	84.00	82.00
11	vannapoeg	49.00	74.00	53.00	81.00	77.00	80.00	78.00
12	vannatütar	42.00	83.00	53.00	74.00	85.00	72.00	85.00
13	õde	77.00	10.00	78.00	52.00	63.00	73.00	83.00
14	poeg	85.00	52.00	82.00	14.00	34.00	80.00	74.00
15	onu	10.00	77.00	39.00	85.00	72.00	85.00	78.00

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Näide 5

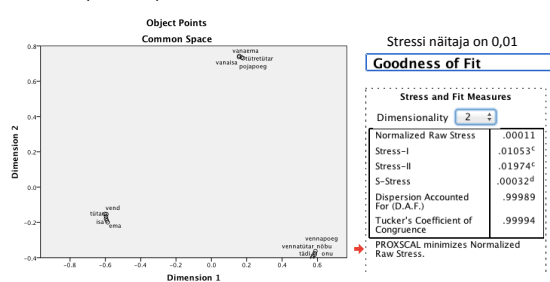
- 1D (Proxcal) – selged kolm gruppi (veresuguluse astme lähedus)



Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Näide 5.

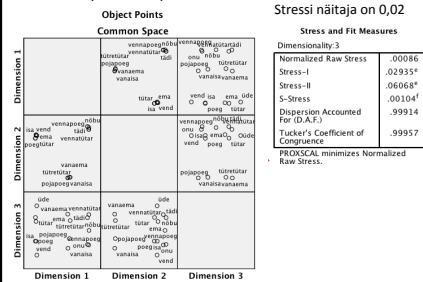
- 2D (Proxcal)



Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Näide 5.

- 3D (Proxcal)



Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Ülesanne 6.

- Kasutades MDS ALSCAL vahendit, leia 1D, 2D ja 3D mudelid.
- Tõlgenda saadud mudeleid ning otsusta, millist mudelit kasutaksid. Põhjenda vastust.

Järgmine andmestik

- Kalevi kommid.sav
- Objektideks on selles andmestikus ...
- Probleemiks saab nende andmete korral ... ning mõelda tuleks andmete
- Kas grupeerime objekte või tunnuseid?

Näide 7.

- Grupeerime tunnuseid
- ALSCAL
- Variables
- Create distances from data
 - Interval (Euclidean distance)
 - Create Distance Matrix (Between variables)
- Model
- Options

Näide 8.

- Grupeerime objekte
- ALSCAL
- Standardiseerime andmed
- Standardiseerimata andmed
- Milline nendest mudelitest grupeerib selgemalt? Millise stressi näitaja parem?
- Kas võime mõlema näite korral öelda, et oleme viiest tunnusest kokku saanud kaks e andmed taandanud kahele dimensioonile?