

Eessõna

Käesolev õppematerjal on mõeldud üliõpilastele, kellel on omandatud statistika põhitõed, kuna multidimensionaalne skaleerimine on üks statistilise analüüsi meetod. Kasu oleks keskkooli matemaatika kursuse eelnevast läbimisest. Antud töös on tahetud selgitada paljude andmete koos analüüsimise põhimõtteid kõige lihtsamal kujul, mis peaks lugejale andma arusaamise selle dissipliini põhilistest kasutusvaldkondadest koos rakendamisoskusega lihtsamates situatsioonides. Käesolev õppevahend ei hõlma kaugeltki kõike MDS-iga seonduvat. Samuti on võimalikult vähe üritatud süveneda valemitesse ja nende tuletamisse.

Eeldusel, et teksti lugeja pole tuttav mitmedimensionaalse analüüsi terminitega on alustuseks selgitatud kõige põhilisemaid mõisteid. Teksti sees käsitletakse ka teisi antud valdkonda kuuluvaid kuid mitte nii määrava tähtsusega termineid.

Kuna mitmedimensionaalses analüüsis üritatakse suurt hulka algandmeid lihtsustatud ja selgemal kujul esitada, siis läheb lihtsutamise käigus alati mingi hulk algandmete variatiivusest kaduma. Sellega seoses selgitatakse kuidas määrata piir milleni on mõtet algandmeid lihtsustada, ilma et liiga palju infot kaduma läheks ning samal ajal oleksid andmed siiski selgemal kujul esitatud.

Sissejuhatus

MDS ühendab endas tervet hulka meetodeid, mis iseloomustavad uuritavate objektide vaheliste suhete struktuuri, võimaldades neid suhteid esitada geomeetrilisel kujul. MDSi meetodid kuuluvad mitme muutujaga andmeanalüüsi üldisemasse kategooriasse. Mitmemõõtmelist skaleerimist saab iseloomustada uuritud suhete üldistuse taseme kaudu. Andmeid suhete üldistamise tasemest saab ühest küljest ära kasutada andmete analüüsis ja teisalt aitab üldistuse tase täpsustada nende suhete geomeetrilise esitluse tüüpi. Ühe objektide paari omavahelist suhet (mida saab esitada läheduse mõõduna või vastupidiselt erinevuse mõõduna) võib vaadelda mitmemõõtmelise skaleerimise sisendandmena. Kõige tähtsamat osa nn. modelleerimises (ehk graafiku koostamises) kujutab kindlat tüüpi maatriksi ruumilise esitluse valik, mis seisneb maatriksis olevate kaugus- või sarnasusmõõtude töötlemises vastava kindlat tüüpi MDS-i algoritmiga.

Mõisted

Mitmemõõtmeline skaleerimine - MDS (ingl. k. *Multidimensional Scaling*) on tehnika, mida kasutatakse mitmemõõtmeliste ehk paljude tunnuste abil kirjeldatud andmete struktuuri võimalikult täpseks ja lihtsaks graafiliseks esitamiseks ühe-, kahe- või kolmemõõtmelisel kujul.

Kaugus ehk erinevus – (ingl.k. *dissimilarity*) on üldine termin, mis viitab objektide vahelisele distantstile, mis on mõõdetud nii, et suuremad väärtused tähistavad objektide suuremat omavahelist kaugust/erinevust.

Lähedus ehk sarnasus - (ingl.k. *similarity*) on üldine termin, mis viitab objektide vahelisele distantstile, mis on mõõdetud nii, et suuremad väärtused tähistavad objektide suuremat omavahelist lähedust/sarnasust.

Koormus - (ingl.k. *stress*) on mõõt, mis kirjeldab lähteandmete (originaalkauguste) ja MDS'i käigus saadud lahenduse (tuletatud kauguste) omavahelist vastavust ehk hinnang sellele, millisel määral on lähteandmed moondunud MDS-i abil lihtsama esitluse leidmise käigus. Mida väiksem koormus, seda parem vastavus on algandmete ja lahendi vahel.

Kasutusvaldkonnad:

Muldidimensionaalset skaleerimist saab kasutada kõikides valdkondades, kus on tegemist suure hulga keeruka struktuuriga andmetega, mida on raske tervikuna mõista. Töödeldes suurt hulka andmeid esitab MDS võimalikult lihtsalt nende andmete üldist ülesehitust näiteks kahemõõtmelise graafikuna.

Järgnevas loetelus on äratoodud mõned levinumad MDSi kasutusvaldkonnad tähestikulises järjekorras koos näidetega vastavatest teadusharudest:

Astronoomia

- magnettormide analüüs

Bioloogia:

- vee kvaliteedi hindamine
- keskkonna mõjude hindamine

Geodeesia:

- topograafiline kaardistamine

Majandus:

- turu-uuringud

Matemaatika :

- faktoriaalanalüüs
- klasteranalüüs

Psühholoogia:

- käitumise uuringud (seksuaal- , virtuaal -,jne. käitumised)

Semiootika:

- kultuuri uuringud (AIDS'i ja teiste haiguste sotsiaal- epiteemioloogia jne.)

Sõjandus

- vaenlase asukoha määramine jne.

Sisendandmed

Sisendandmeteks on MDSi puhul uuritud objektide vahelised suhted, mis on esitatud ruutmaatriksisse organiseeritud kaugus- või sarnasusmõõtudena. Sarnasusmaatriksi puhul näitavad suuremad numbrid suuremat sarnasust ja kaugusmaatriksi puhul suuremat erinevust.

		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	

Näide: Linnade vahelised kaugused mõõdetuna arvestades kattega riigimaanteid

Reaalses andmeanalüüsi situatsioonis võib selline maatriks tekkida otsese/objektiivse mõõtmise tulemusena (vt näide ülal) või subjektiivse hindamisena. Viimasel juhul palutakse vastajatel hinnata etteantud objektide (nt riikide, teatud kaubamärkide, jms) sarnasust või erinevust lähtuvalt vastaja sisemisest tunnetusest.

Olenevalt skaala tüübist, mille põhjal sarnasust hinnatakse, võib siis iga maatriksi ruut (x_{ij}) tähistada nt.

konkreetses mõõtmistulemuse väärtust, võrdluste keskmist tulemust, erinevate/sarnaste elementide absoluutarvu või suhtarvu, mis näitab, mitu inimest liigitasid objektid i ja j samasse kategooriasse.

Saadud väärtuse järgi võib näiteks hinnata sihtgrupi arusaamu objektide sarnasusest – mida suurem number seda sarnasem. Graafikul esitades oleksid sarnased objektid lähestikku ja erinevad teineteisest kaugel.

Teisel juhul on algandemeteks tüüpiline mitmemõõtmeline andmetabel, milles olevate tunnuste väärtuste põhjal soovitakse tegelikud kaugus- või sarnasusmõõdud arvutada (st sarnasusmaatriks koostada).

Kui esimesel juhul on tavaliselt mitmemõõtmelise skaleerimise eesmärgiks teada saada nn peidetud mõõtmised, millest lähtuvalt teatud isikud või grupid objektide sarnasust hindavad ja/või kirjeldada ning selgitada sarnasusmustrite struktuuri ja võimalikku tekkeloogikat, siis teisel juhul on tavaliselt eesmärgiks sarnasuse hindamise aluseks olevate paljude mõõtmete taandamine väiksemalt mõõtmete hulgale, mis võimaldaks paremini ja selgemini objekte lähtuvalt nende sarnasusest klassifitseerida.

MDSi etapid

MDSi mudelid võimaldavad kasutada erinevaid algoritme vahekauguste geomeetrilise esitluse saavutamiseks. Mudeleid saab liigitada lähtuvalt sellest milliste tunnuse e skaala tüüpide jaoks nad on sobivad. Osad mudelid sisaldavad endas ühtlasi ka üldistuste tegemise ja staistilise olulisuse testimise võimalusi.

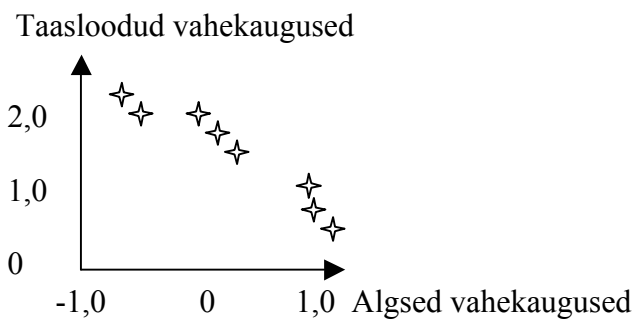
Lihtsustatud etappide seletus:

1. Koguda algandmed kasutades valitud skaalasid ja sisestada need kas tavalise andmestiku või kauguste maatriksina.
2. Kui algandmed ei ole kauguste maatriksi kujul, siis arvuta objektide omavahelised kaugused (a_{11}) lähtudes geomeetria reeglitest ning esita nad maatriksina.
3. Vali sobiv mitmedimensionaalse skaleerimise mudel ning soovitud dimensioonide arv (võib valida korraga mitu varianti, et siis tekkinud mudeleid võrrelda) ja rakenda seda tuletatud koordinaatide ja tuletatud kauguste arvutamiseks.
4. Võrdle saadud kauguste maatriksit esialgse kauguste maatriksiga koormuse funktsiooni (*stress*) kaudu. Mida väiksem on koormuse väärtus seda väiksem on maatriksite erinevus (st paremini sobiv on mudel).
5. Leia tekkinud dimensioonidele/mõõtmetele tõlgendus.
6. Vajadusel/soovi_korral korda samme 3-5 muutes arvutuste aluseks olevaid parameetreid
7. Vali lähtuvalt koormuse näitajast ning tõlgenduse sisukusest sobivaim lahend

8. Koormus

Koormus on hinnang, kui hästi või halvasti antud graafik loob uuesti uuritud kauguste maatriksit. Mida väiksem on koormuse väärtus, seda parem on MDS'iga töödeldud kauguste maatriksi sobivus uuritud esialgsete kauguste maatriksiga.

- **Shepard'i diagrammiks** nimetatakse graafikuid taasloodud vahekauguste kohta, mida on võrreldud algsete sisestatud andmetega (graafiline koormuse kujutamine). Graafik näitab taasloodud vahekaugusi vertikaalsel teljel ja algseid vahekaugusi horisontaal ehk x- teljel (mõõdetud läheduste puhul peaks tulemuseks olema punktide rea negatiivne kalle ning erinevuste puhul positiivne kalle). Kui koormus võrdub nulliga siis peaksid uuritavaid objekte iseloomustavad punktid (esialgsed vahekaugused x_{ij} , MDSiga taasloodud vahekaugused d_{ij} kui ka funktsioonid nendest vahekaugustest ($f(x_{ij})$ ja $f(d_{ij})$) asetsema täpselt samas punktis. Seega mida suurem on ühte objekti iseloomustavate punktide hajuvus, seda suurem on koormus.



Mudeli „headust” e kauguste taasesitamise täpsust iseloomustava koormusnäitaja tõlgendamisel võib lähtuda järgmistest kriteeriumitest:

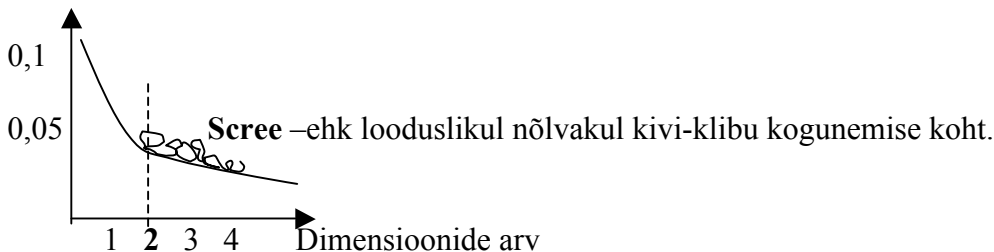
Stress (type I)		Hinnang vastvusele
Suurem kui	0,20	halb
Väikesem kui	0,05	hea
Väiksem kui	0,00...	väga hea

Dimesioonide (mõõtmete) arv

Põhimõtteliselt võiks analüüsida näiteks mingite objektide hulga 50 erinevat omadust korraga, kuid sellele seab inimmõistus piirid (kuna iga omadus tähendab uut mõõdet) ja maksimaalselt saab visuaalselt esitada vaid kolmemõõtmelist graafikut. Uurides 50 omadust korraga võiks põhimõtteliselt andmeid lihtsustada kahemõõtmeliseks, kuid sellisel juhul tekib arvatavasti väga kõrge “koormuse” väärtus. Sellisel juhul on kaks valikut - kas loobuda MDSi meetodist või suurendada dimensioonide arvu kolmeni ning rahulduda mitte väga hästi sobiva mudeliga eeldusel, et sellel on siiski andmete struktuuri kirjeldamise seisukohast sisuline väärtus.

- **Scree katse.** Tavaline moodus mõõtmete arvu määramiseks on kasutada koormuse ja erinevate mõõtmete vahelist graafikut. Graafikult peab leidma subjektiivselt koha, kus vasakult paremale allalaskuv joon hakkab parempoolse joonekalde väärtusele lähenema. (Scree- inglise keeles – “kivi-klibu” tuleb sellest, et kivise nõlva kallaku sellele piirile koguneb klibu.)

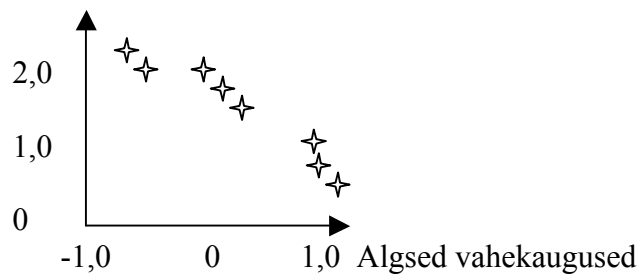
Koormus



Teine moodus mõõtmete arvu määramiseks on selgitada välja kõige selgem lõplik asend. Mõningatel kordadel ei ole punktide asend selge, vaid moodustab nn. juhusliku pilve, mille puhul ei ole otsest ja selget tõlgendamise võimalust. Sellisel juhul peab proovima mõõtmeid ära võtta või juurde panna.

MDS- i telgesid saab keerata igasse suunda ja see on viimane analüüsi etapp. Tõlgendamise hõlbustamiseks tuleb andmematriksi MDSiga töötlemise tulemuste põhjal luua objektidest kahemõõtmelisi graafikuid (kolmemõõtmelised kipuvad olema keerukamad). Oluliseimate mõõtmete otsinguil tuleks tähelepanu pöörata punktide hulgas tekkivatele kujunditele nagu ringid, kõverikud jne.

Taasloodud vahekaugused



Mudeli „headust” e kauguste taasesitamise täpsust iseloomustava koormusnäitaja tõlgendamisel võib lähtuda järgmistest kriteeriumitest:

Stress (type I)		Hinnang vastvusele
Suurem kui	0,20	halb
Väikesem kui	0,05	hea
Väiksem kui	0,00...	väga hea

Koormus

