TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
MATEMAATILISE STATISTIKA INSTITUUT

Gertis Aru

Korrespondentsanalüüs ja andmete dubleerimine

Bakalaureusetöö

Juhendaja:

prof. Kalev Pärna

TARTU

2013

Sisukord

Sissejuhatus

Korrespondentsanalüüsi saab kasutada väga hästi risttabeliga, kus on 2 rida või 2 veergu, saades numbrilised väärtused nii rea, kui ka tulpade kategooriatele. Need väärtused (skoorid) saadakse nii, et need kirjeldaks kahe tunnusevahelist seost nii palju kui võimalik. Enamasti on rea ja tulba kategooriad esitatud kahemõõtmelisel graafikul, kus asuvad vastavate skooride paarid. Selline esitlus annab lugejale ülevaate rea- ja veerukategooriate erinevustest ja sarnasustest. See meetod on väga sarnane peakomponentide analüüsiga, mida saab kasutada selleks, et selgitada välja dimensioonid, mis kirjeldavad andmestikku kõige paremini. Korrespondentsanalüüsi saab kasutada ka väga suurte andmestike korral, kus nii ridade kui ka veergude arv on väga suur. Korrespondentsanalüüsi saab läbi viia mitmete statistikaprogrammide abil, nagu näiteks SAS, SPSS, BMDP jne. Meie kasutame siin töös aga rakendustarkvara SAS.

Töö on jaotatud kahte suurde peatükki. Esimeses osas on autori eesmärk anda teoreetiline ülevaade korrespondentsanalüüsi matemaatilisest meetodist. Töö teises pooles on korrespondentsanalüüsi rakendamine reaalsete vaatlusandmete põhjal. Autor on lisaks viinud läbi ka andmete dubleerimise, et vaatlusandmeid põhjalikumalt analüüsida.

Autor tänab professor Kalev Pärnat rohkete täienduste ja paranduste ning mitmete töö struktuuri puudutavate ideede eest.

1. Korrespondentsanalüüsi meetod

Antud peatükis anname ülevaate korrespondentsanalüüsi matemaatilise meetodi kohta. Kasutame samu tähiseid, mida on kasutatud oma töös Pärna(1993).

1.1. Rea- ja veeruprofiilid. Pilv

Olgu meil n vaatlust (isikut, objekti), mis on jaotatud tunnuste *A* ja *B* järgi. Oletame, et juhuslikul suurusel *A* on I kategooriat $A_1, A_2, \dots, A_I$ ja juhuslikul suurusel *B* on J kategooriat $B_1, B_2, \dots, B_J$. Olgu $n_{i.}$ vaatluste arv, kus tunnuse väärtus on A_i ja sarnaselt $n_{.j}$ olgu vaatluste arv, kus tunnuse väärtus on B_j. Vaatluste arv, kus on korraga mõlemad väärtused A_i ja B_j tähistatakse n_{ij}. Seega saab andmestiku esitada (IxJ)-risttabelina kujul $N=(n_{ij})$, kus rea-, veeru- ja kogusummad avalduvad järgnevalt:

$$n_{i.} = \sum_j n_{ij},$$

$$n_{.j} = \sum_i n_{ij},$$

$$n = \sum_i \sum_j n_{ij}.$$

Korrespondentsanalüüsis huvitavad meid kõige enam tinglikud jaotused risttabelis *N*. Järgnevalt tähistame suhtelised sagedused (tõenäosused) tabelis *N*:

$$f_{ij} = n_{ij}/n \ ,$$

$$f_i = n_{i.}/n \ ,$$

$$f_j = n_{.j}/n \ ,$$

$$f_j^i = n_{ij}/n_{i.} = f_{ij}/f_i \ ,$$

$$f_i^j = n_{ij}/n_{.j} = f_{ij}/f_j \ .$$

Vektor $f_B^i = (f_1^i, f_2^i, \dots, f_J^i)^T$ tähistab veeru tunnuse *B* tinglikku jaotust tingimusel, et reatunnus $A=A_i$. Edaspidi nimetame vektorit f_B^i *reaprofiiliks* (rea i jaoks). Sarnaselt, vektor

$f_A^j = (f_1^j, f_2^j, \dots, f_I^j)^T$ tähistab reatunnuse *A* tinglikku jaotust tingimusel, et veerutunnus $B=B_j$. Vektorit f_A^j nimetatakse *veeruprofiiliks* (veeru j jaoks).

Paneme tähele, et I reaprofiili $f_B^1, \dots, f_B^I$ määravad I punkti J-dimensionaalses eukleidilises ruumis. Me nimetame neid I punkte *pilveks* ruumis R^J. Tähistame rea- ja veeruprofiilide pilve järgnevalt:

$$N_B(A) = \{f_B^i | i = 1, \dots, I\}.$$

$$N_A(B) = \{f_A^j | j = 1, \dots, J\}.$$

1.2. Punkti mass

Korrespondentsanalüüsis on väga tähtis aru saada, et igal punktil f_B^i pilves $N_B(A)$ on oma mass, mis on defineeritud marginaaltõenäosusega f_i (rea i korral). Seega punktipilv on kaalutud punktide kooslus. See lubab meil defineerida pilve keskpunkti, mis on pilve masskese. Ehk teisisõnu, see on pilve kõikide elementide kaalutud keskmine. Tähistame pilve $N_B(A)$ masskeskme f_B, mis avaldub järgneval kujul:

$$f_B = \sum_i f_i \cdot f_B^i$$

See on omakorda võrdne suuruse *B* marginaaljaotusega:

$$f_B = (f_1, f_2, \dots, f_J)^T.$$

Selgitame nüüd pilve dimensionaalsuse probleemi. Teame, et kõik punktid pilves $N_B(A)$ on J-dimensionaalse ruumi elemendid. Kuna tinglike jaotuste tõttu, vektorite tõenäosused summeeruvad üheks, siis näeme, et profiilid asuvad tegelikult (J-1)-mõõtmelises alamruumis. Teisest küljest, saab iga I punktilist hulka vaadelda (I-1)-mõõtmelises alamruumis. Seega me ei vaja pilve kujutamiseks rohkem kui $min\{J-1, I-1\}$ dimensiooni. Olenevalt andmestikust võib tegelik dimensioon olla isegi väiksem. Näiteks, me kaotaksime veel ühe dimensiooni, kui tabel *N* sisaldab kahte võrdset rida. Sellisel juhul tekiks meil kaks identset reaprofiili. Kokkuvõttes, „õige“ dimensioon on defineeritud maatriksi *N* astaku kaudu järgnevalt:

$$K = rank(N) - 1 \leq \min\{J - 1, I - 1\}\,. \quad (1)$$

Samamoodi saame kirjeldada ka nö duaalset pilve. Selleks on veeruprofiilide pilv, mis koosneb J punktist I-dimensionaalses ruumis. Tähistame seda pilve järgnevalt:

$$N_A(B) = \{f_A^j | j = 1, \dots, J\}.$$

Pilve $N_A(B)$ elementidel on massid, mis vastavad marginaaltõenäosustele f_j. Ka siin pilves on keskpunkt f_A, milleks on kõikide elementide kaalutud keskmine. Seekord on aga keskpunkt võrdne tunnuse *B* marginaaljaotusega:

$$f_A = (f_1, f_2, \dots, f_I).$$

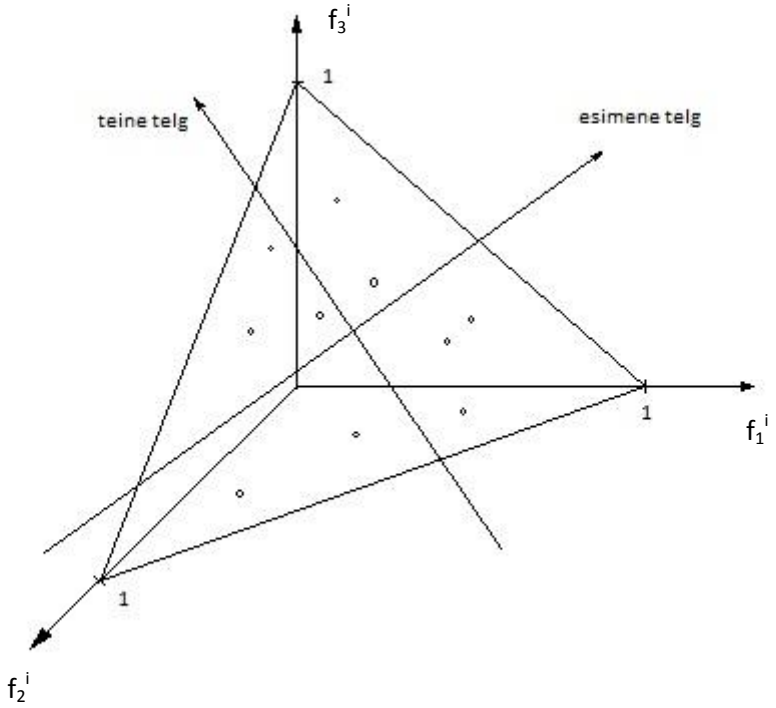
1.3. Korrespondentsanalüüsi eesmärk

Formuleerime nüüd korrespondentsanalüüsi eesmärgi, nii nagu selle on välja toonud oma töös Greenacre(1993). Geomeetrilistes terminites on meie eesmärk leida madala dimensiooniga alamruum, mis kirjeldaks pilve kõiki punkte võimalikult täpselt. Täpsemalt, tuleb leida kaks peatelge, mis kirjeldaks meie punktipilve võimalikult täpselt. Punktide lähedust silmas pidades oleks vaja, et punktid, millel on suuremad massid, peaksid asetsema alamruumile väga lähedal. Samal ajal punktid, millel on väiksemad massid, võivad asuda otsitavast alamruumist kaugemal.

Et ideed lugejale selgemaks teha, siis oletame, et meil on andmestik, mida võib kujutada 10x3 risttabelina. Pärast kümne reaprofiili arvutamist saame, et tegelikult asetsevad need punktid 2-dimensionaalses ruumis, kus:

$$f_1^i + f_2^i + f_3^i = 1,\ \forall\, f_j^i \geq 0,$$

Kogu andmestik on kujutatud Joonisel 1.



Joonis 1. Kõik kümme reaprofiili asuvad kolmnurga peal

Pöördume nüüd tagasi esialgse pilve $N_B(A)$ juurde, kus asetsevad reaprofiilid, ning defineerime elemendi $f_B{}^i$ kauguse pilve keskpunktist järgnevalt:

$$d^2(i) = \sum_j (f_j^i - f_j)^2 / f_j \ . \qquad (2)$$

Seda avaldist kutsutakse ka hii-ruut kauguseks, sest see on kooskõlas hii-ruut statistikuga, mida kasutatakse tingliku jaotuse f_j^i võrdlemiseks jaotusega f_j. Kaugus (2) erineb tavalisest eukleidilisest kaugusest kahe punkti (profiili ja keskpunkti) vahel, sest see valem sisaldab endas normeerivat faktorit f_j. Kui me aga muudame seda ruumi, tehes selle j-nda baasivektori $\sqrt{f_j}$ korda pikemaks, ehk mis on sama, kui vähendada j-ndat koordinaati $\sqrt{f_j}$ korda, siis valem (2) on selles muudetud ruumis sama, mis eukleidiline kaugus. Seega me võime käsitleme valemit (2) kui kaalutud eukleidilist kaugust.

1.4. Inerts

Üheks tähtsamaks mõisteks korrespondentsanalüüsis on inerts. Inerts mõõdab, kui palju profiilid on hajutatud ümber masskeskme. Pilve $N_B(A)$ inerts on defineeritud kui I profiilipunkti kaalutud keskmine kaugus pilve masskeskmest:

$$in(A) = \sum_i f_i \cdot d^2(i).$$

Inerts on väga sarnane hii-ruut statistikuga, mida kasutatakse kahemõõtmelise risttabeli sõltumatuse testimiseks. Näitame ka valemina seost nende kahe vahel:

$$in(A) = \sum_i \sum_j \frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n})^2}{n_{i.} n_{.j}} = \chi^2/\mathrm{n}\ . \qquad (3)$$

Duaalse pilve $N_A(B)$ inerts on defineeritud järgnevalt:

$$in(B) = \sum_j f_j \cdot d^2(j),$$

kus

$$d^2(j) = \sum_i (f_i^j - f_i)^2 / f_i$$

on veeruprofiilide f^j kaalutud eukleidiline kaugus pilve masskeskmest f_B. Valem (3) kehtib ka duaalse pilve korral:

$$in(B) = \chi^2/\mathrm{n},$$

Tähistame mõlemad inertsid järgnevalt:

$$\lambda = in(A) = in(B) = \chi^2/n.$$

Madala dimensiooniga alamruum, mis kirjeldab pilve punkte võimalikult täpselt, on leitav inertsi *peatelgedega*. Peateljed on K vektorit, mis on rakendatud pilve masskeskmele ja näitavad suurimate inertside suunda. Seejuures iga järgnev telg on kõigi eelnevatega ortogonaalne. Et saada madala dimensiooniga ruum, siis on vaja vaid esimesi (näiteks K^*) peatelgi lootuses, et need K^* dimensiooni kirjeldavad meie andmestikku võimalikult hästi.

Väga tihti kasutatakse praktikas andmestiku kujutamiseks vaid kahte esimest dimensiooni, kus 2-dimensionaalsel graafikul on esitatud rea- ja veeruprofiilid.

1.5. Singulaarse lahutuse kasutamine korrespondentsanalüüsis

Punktipilvi hästi kirjeldava madala dimensiooniga ruumi leidmiseks kasutatakse singulaarset lahutust (*singular value decomposition*). Selle tutvustamiseks võtame kasutusele täiendavad tähistused. Olgu *R* ja *C* andmetabeli *N* rea- ja veeru summade diagonaalmaatriksid vastavalt:

$$R = diag(n_{1.}, \dots, n_{I.})$$

ja

$$C = diag(n_{.1}, \dots, n_{.J}).$$

Olgu m_1,..., m_K reaprofiilide pilve inertsi peateljed. Praktikas on märksa olulisem teada punktide koordinaate nende baasivektorite suhtes. Olgu x_{ik} i-ndale reaprofiilile vastav koordinaat k-nda peatelje sihis. Edasipidi kutsume seda *peakoordinaadiks*. Sellisel juhul on i-s reaprofiil kirjeldatav K-vektori kaudu:

$$x_i = (x_{i1}, x_{i2}, \dots, x_{iK}),$$

kus K oli defineeritud seosega (1) ja kogu pilve reaprofiilid on kirjeldatavad (IxK)-maatriksi *X* kaudu:

$$X = \begin{pmatrix} x_1 \\ \dots \\ x_I \end{pmatrix}.$$

Singulaarse lahutuse teooriast tuleneb, et maatriksi *X* kõik K veergu peavad olema (IxI)-maatriksi $R^{-1}NC^{-1}N^T$omavektorid, mis vastavad mittenullilistele omaväärtustele $\lambda_1, \dots, \lambda_K$. See omakorda tähendab, et maatriks *X* rahuldab järgnevat võrrandit:

$$(R^{-1}NC^{-1}N^T)X = XD_\lambda, \qquad (4)$$

kus D_λ on KxK omaväärtuste diagonaalmaatriks, $D_\lambda = diag(\lambda_1, \dots, \lambda_K)$, eeldades järjestust $\lambda_2 \geq \cdots \geq \lambda_K > 0$. Lugeja võib tähele panna, et mittenulliliste omaväärtuste arv on $rank(N) = K + 1$. Seejuures, me ei kasutanud suurimat omaväärtust $\lambda_0 = 1$ ja vastavat

„triviaalset“ omavektorit, mis koosneb numbritest 1, sest see ei paku meie edasises arutelus huvi. Geomeetriliselt tähendab triviaalse omaväärtuse ja omavektori mittekasutamine seda, et oma analüüsis paigutame peatelgede alguspunkti pilve masskeskmesse.

Et saada ühest lahendust omaväärtuste probleemile (4), on oluline täpsustada omavektorite maatriksi X veergude pikkused. Oleks mõistlik kasutada standardiseerimist.

$$\frac{1}{n}X^T RX = D_\lambda,$$

mis määrab, et punkti koordinaatide ruutude kaalutud summa (ehk kaalutud dispersioon või inerts) piki pilve k-ndat peatelge on võrdne omaväärtusega λ_k. Seda omaväärtust nimetame k-ndaks *peainertsiks*.

Asjaolu, et reaprofiilide masskese on peatelgede alguspunkt, saab esitada järgnevalt:

$$\sum_i n_{i.} x_i = 0. \qquad (5)$$

Sellest tulenevalt saame, et iga dimensiooni korral peavad mõned profiili koordinaadid olema negatiivsed.

Rakendades korrespondentsanalüüsi on väga tavaline kasutada vaid kahte esimest inertsitelge. See tähendab, et enamasti võetakse vaid kaks esimest maatriksi *X* veergu, et saada iga reaprofiili esimene ja teine koordinaat. Neid koordinaate kasutatakse I reaprofiili kujutamiseks tasandil.

Kui meil on aga tegu duaalse pilvega, mis sisaldab endas veeruprofiile, siis tähistame inertsi peatelgi $l_1,\ldots,l_K$ ja vastavalt j-nda veeruprofiili koordinaate $y_{j1},\ldots,y_{jK.}$ Koordinaadid saab esitada (JxK)-tabelina *Y*. Singulaarse lahutuse meetod eeldab, et *Y* peab rahuldama järgnevat võrrandit

$$(C^{-1}N^T R^{-1}N)Y = YD_\lambda,$$

kus D_λ on sama maatriks nagu valemis (4). Nüüd oleks sobiv standardiseerimine järgmine:

$$\frac{1}{n}Y^T CY = D_\lambda,$$

mis annab meile peainertsi koos iga reaprofiilide pilve teljega. Erinevalt (5) on meil nüüd maatriksi *Y* read elementidega y_j, mis peavad rahuldama järgnevat võrrandit:

$$\sum_j n_{.j} y_j = 0.$$

Peainertside võrdsus mõlemas pilves lubab meil ühendada kaks eraldiseisvat graafilist esitust üheks, kus on I reaprofiili punkti ja J veeruprofiili punkti. Greenacre'i (1984) kohaselt peaksime siiski hoiduma erinevates pilvedes asetsevate punktidevahelise kauguse esitamisest, sest selliseid kauguseid pole otseselt defineeritud. Selle põhjuseks on asjaolu, et me kasutame kahte erinevat baasi kummagi pilve jaoks ja seetõttu pole koordinaadid otseselt võrreldavad. Siiski on olemas seos kahe paari peakoordinaatide jaoks:

$$YD_\rho = C^{-1}N^T X, \tag{6}$$

$$XD_\rho = R^{-1}NY, \tag{7}$$

kus $D_\rho = {D_\lambda}^{1/2}$, mis on diagonaalmaatriks, kus $\rho_k = \sqrt{\lambda_k}$ asetsevad peadiagonaalil. Seosed (6) ja (7) võib esitada ka elemendiviisiliselt:

$$\rho_k y_{jk} = \frac{1}{n_{.j}} \sum_i n_{ij} x_{ik},$$

$$\rho_k x_{ik} = \frac{1}{n_{i.}} \sum_j n_{ij} y_{jk}.$$

Esimene võrrand näitab meile, et iga veerukoordinaat on reakoordinaatide kaalutud keskmine (konstandi ρ_k täpsusega), kus kaaludeks on sagedused n_{ij}. Teine võrrand on analoogiline. Sellist seost koordinaatide vahel kutsutakse vastastikuseks keskmistamiseks ning seda kasutatakse peakoordinaatide praktiliseks leidmiseks tihti, eriti kui tegu on väga suure andmestikuga. Mõnikord kasutatakse seda terminit isegi tähistamaks korrespondentsanalüüsi.

Viimaks anname ülevaate omaväärtuste $\lambda_1, \dots, \lambda_K$ tähtsusest, näidates, et nad on tegelikult koguinertsi λ komponendid. Nagu eelnevast mäletame, näitas koguinerts seda, kui palju profiilid on hajutatud ümber pilve masskeskme. Selle mõõdu võib aga jagada K komponendiks järgnevalt:

$$\chi^2 = n\left(\sum_i \sum_j \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1\right)$$

$$= n(tr(R^{-1}NC^{-1}N^T) - 1)$$

$$= n(\lambda_1 + \lambda_2 + \cdots + \lambda_K),$$

sest maatriksi jälg on võrdne omaväärtuste λ_k summaga. Seega näeme, et:

$$in(A) = in(B) = \lambda_1 + \lambda_2 + \cdots + \lambda_K$$

See tähendab, et koguinerts on jaotatud K peatelje vahel, kus esimene telg kirjeldab kõige rohkem jne. See on koguinertsi lahutuse valem, mis ühtlasi on hii-ruut statistiku lahutuse valem risttabelis. Suhtelisi inertse λ_K/λ (mõõdetakse protsentides) kasutatakse erinevate peatelgede tähtsuse väljendamiseks.

2. Korrespondentsanalüüsi rakendamine

2.1. Standardne rakendus

Uurisime 2012. aasta Londoni Olümpiamängude kümnevõistluse tulemusi. Kümnevõistluse kõik alad lõpetas 26 võistlejat. Andmestiku read vastavad 26 võistlejale ja tulpadeks on 10-l alal saavutatud punktid ning kogusumma. Võistlejad on andmestikku lisatud paremusjärjestuses, kus esimesena on lisatud võitja ja viimasena 26. koha omanik. Alad on tähistatud nende toimumisjärjekorras, kusjuures alad a1-a5 toimuvad kümnevõistluses esimesel päeval ja alad a6-a10 teisel päeval:

a1 – 100m jooks
a2 – kaugushüpe
a3 – kuulitõuge
a4 – kõrgushupe
a5 – 400m jooks
a6 – 110m tõkkejooks
a7 – kettaheide
a8 – teivashüpe
a9 – odavise
a10 – 1500m jooks

Polnud mõistlik kasutada oma korrespondentsanalüüsis tunnust „kogusumma“, sest selle lisamine ei anna meile uut informatsiooni.

Järgneva korrespondentsanalüüsiga tahame uurida, kas on mingisuguseid seoseid võistlejate profiilide vahel. Näiteks, kas mõned profiilid moodustavad mingisuguse klastri või on näiteks teistest väga eraldi seisvaid võistlejaid. Lisaks proovime leida seoseid võistlejate ja alade vahel. Näiteks, kas mingi võisteja eristus mingi ala tulemuse poolest teistest märgatavalt. Et saaks uurida vastavaid seoseid, selleks viime programmis SAS läbi protseduuri CORRESP.

Tabel 1. Inertsid ja Hii-ruut statistiku lahutus

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	7 14 21 28 35
0.04481	0.00201	419.37	34.13	34.13	**********
0.03762	0.00142	295.63	24.06	58.20	**********
0.03082	0.00095	198.34	16.14	74.34	**********
0.02167	0.00047	98.11	7.99	82.33	******
0.01933	0.00037	78.03	6.35	88.68	*****
0.01640	0.00027	56.14	4.57	93.25	***
0.01516	0.00023	48.02	3.91	97.15	***
0.01078	0.00012	24.29	1.98	99.13	*
0.00715	0.00005	10.67	0.87	100.00	*
Total	0.00588	1228.60	100.00(Degrees of Freedom = 225)		

Tabel 1 koosneb viiest tulbast. Esimenes tulbas (*Singular Value)* on omaväärtused, mis leitakse siis $\rho_k = \sqrt{\lambda_k}$. Teises tulbas on peainertsid λ_k. Kolmandas tulbas on χ^2-statistiku väärtused. Neljas tulp näitab, mitu protsenti koguinertsist antud telg kirjeldab. Viiendas tulbas on eelnevad protsendid kokku liidetud. Meie näites on peaaegu 75% kirjeldatud esimese kolme telje poolt. Seega on meie andmestik oma põhiolemuselt kolme-mõõtmeline. Kuid kuna me tahame anda graafilist ülevaadet, siis huvitavad meid eeskätt kahe esimese telje väärtused. Need kirjeldavad kogu inertsist 58%.

Tabel 2. Reaprofiilide koordinaadid

	Dim1	Dim2
EATON	-0.0039	-0.0483
HARDEE	-0.0137	0.0176
SUAREZ	0.0873	0.0296
VANALPHE	0.0332	0.0157
WARNER	0.0196	0.0111
FREIMUTH	-0.0352	0.0018
KASYANOV	-0.0252	0.0122
SVIRIDOV	0.0436	0.0410
COERTZEN	0.0461	0.0197
BEHRENBR	0.0155	0.0116
SINTNICO	0.0316	-0.1117
NEWDICK	0.0107	0.0125
BARROILH	-0.0074	-0.0862
GARCIA	-0.0021	0.0014
MAYER	0.0739	-0.0053
SHKURENE	-0.0068	-0.0541
MIKHAN	-0.0283	0.0238
KARPOV	-0.1003	-0.0447
DEARAUJO	-0.0253	0.0014
USHIRO	0.0569	0.0049
VOS	0.0062	0.0109
ERINS	0.0169	0.0209
ADDY	-0.0932	0.0410
SZABO	-0.0092	0.0129
DRAUDVIL	-0.0735	0.0464
ARTIKOV	-0.0374	0.0205

Tabel 2 annab meile numbrilised väärtused x_{i1}, x_{i2} – need on reaprofiilide esimesed kaks koordinaati. Need on koordinaadid, mis näitavad, kus mingi indiviid graafikul asub. Nagu näha, siis näiteks Suarezel on esimesel teljel kõige suurem koordinaadi väärtus, samas aga Karpovil on see väärtus negatiivselt kõige suurem. Teisel teljel on suurim koordinaat Draudvilal ja väikseim koordinaat võistlejal Sintnicolaas. Seejuures nende koordinaatide kaalutud keskmine üle kõikide indiviidide on null.

Tabel 3. Reaprofiilide statistikud

	Quality	Mass	Inertia
EATON	0.4445	0.0425	0.0381
HARDEE	0.2336	0.0415	0.0150
SUAREZ	0.7370	0.0408	0.0799
VANALPHE	0.3235	0.0404	0.0286
WARNER	0.2123	0.0404	0.0165
FREIMUTH	0.3444	0.0398	0.0244
KASYANOV	0.2396	0.0397	0.0221
SVIRIDOV	0.6600	0.0394	0.0363
COERTZEN	0.5562	0.0391	0.0300
BEHRENBR	0.1592	0.0389	0.0156
SINTNICO	0.8181	0.0385	0.1077
NEWDICK	0.1502	0.0382	0.0117
BARROILH	0.6385	0.0382	0.0761
GARCIA	0.0024	0.0381	0.0171
MAYER	0.6999	0.0381	0.0508
SHKURENE	0.5671	0.0381	0.0339
MIKHAN	0.3590	0.0380	0.0246
KARPOV	0.7400	0.0380	0.1051
DEARAUJO	0.1517	0.0376	0.0270
USHIRO	0.4468	0.0375	0.0466
VOS	0.1083	0.0374	0.0092
ERINS	0.3583	0.0366	0.0126
ADDY	0.9143	0.0363	0.0700
SZABO	0.1874	0.0363	0.0083
DRAUDVIL	0.8585	0.0362	0.0541
ARTIKOV	0.2769	0.0345	0.0385

Tabel 3 näitab näitab esmalt kahe-dimensioonilisel graafikul kujutatud reaprofiilide kvaliteeti. Rea i jaoks on see leitud järgnevalt: $(x_{i1}^2 + x_{i2}^2)/d^2(i)$, kus murru nimetaja näitab i-nda punkti kaugust keskpunktist ja murru lugeja on sama profiilipunkti projektsiooni kaugust keskpunktist. Meie näites on võistlejal Addy kvaliteet üle 0,9, mis tähendab, et see punkt praktiliselt ei vaja kolmandat dimensiooni täpsemaks kirjeldamiseks. Nagu näha, siis De Araujo ja Behrenbruchi kvaliteet on üsna nullilähedane. See tähendab, et neid kirjeldavad

esimesed kaks telge üsna halvasti. Geomeetriliselt tähendab see, et need punktid vajavad piisavaks kirjelduseks kolmandat või ka rohkem telgi.

Järgnev tulp *Masses* iseloomustab võistlejate marginaaltõenäosusi p_i meie tabelis. Need summeeruvad ülevalt alla kokku üheks. Kolmas tulp *Inertia* näitab, kui palju antud reaprofiil annab juurde koguinertsile. Reainerts on $p_i \cdot d^2(i)/\lambda$ ning seega kogu tulba summa on 1. Täheldame, et kõige suuremad inertsid on võistlejatel Karpov ja Sintnicolaas. See tähendab, et nende reaprofiilid on rohkem erinevad keskpunktist ning see teeb väärtuse $d^2(i)$ suureks.

Tabel 4. Reaprofiilide panus telgede loomisel

	Dim1	Dim2
EATON	0.0003	0.0700
HARDEE	0.0039	0.0091
SUAREZ	0.1547	0.0252
VANALPHE	0.0222	0.0070
WARNER	0.0078	0.0035
FREIMUTH	0.0246	0.0001
KASYANOV	0.0126	0.0042
SVIRIDOV	0.0373	0.0467
COERTZEN	0.0414	0.0107
BEHRENBR	0.0047	0.0037
SINTNICO	0.0192	0.3389
NEWDICK	0.0022	0.0042
BARROILH	0.0010	0.2005
GARCIA	0.0001	0.0001
MAYER	0.1036	0.0008
SHKURENE	0.0009	0.0787
MIKHAN	0.0151	0.0152
KARPOV	0.1900	0.0536
DEARAUJO	0.0120	0.0001
USHIRO	0.0606	0.0006
VOS	0.0007	0.0031
ERINS	0.0052	0.0113
ADDY	0.1572	0.0431
SZABO	0.0015	0.0043
DRAUDVIL	0.0973	0.0551
ARTIKOV	0.0241	0.0102

Tabel 4 iseloomustab, kui palju annab antud punkt teljele juurde koordinaadi viisiliselt, ehk mida suurem on tulbas number, seda suurema kaaluga on antud punkt telgede loomisel. Tabelist on näha, et esimesele teljele annavad suurima panuse kümnevõistlejad Suarez, Karpov ja Addy. See-eest teisele teljele annavad kõige rohkem juurde Sintnicolaas ja Barrouilhet.

Tabel 5. Millised võistlejad iseloomustavad kõige paremini telgi

	Dim1	Dim2	Best
EATON	0	2	2
HARDEE	0	0	2
SUAREZ	1	0	1
VANALPHE	0	0	1
WARNER	0	0	1
FREIMUTH	0	0	1
KASYANOV	0	0	1
SVIRIDOV	0	2	2
COERTZEN	1	0	1
BEHRENBR	0	0	1
SINTNICO	0	2	2
NEWDICK	0	0	2
BARROILH	0	2	2
GARCIA	0	0	1
MAYER	1	0	1
SHKURENE	0	2	2
MIKHAN	0	0	2
KARPOV	1	1	1
DEARAUJO	0	0	1
USHIRO	1	0	1
VOS	0	0	2
ERINS	0	0	2
ADDY	1	0	1
SZABO	0	0	2
DRAUDVIL	1	1	1
ARTIKOV	0	0	1

Tabel 5 aitab tõlgendada, millised võistlejad kirjeldavad kõige paremalt telgi. Tabelis on kolm tulpa, millest viimane on *Best*. Tulp *Best* näitab, kumb koordinaat (1 või 2) eelmises tabelis oli suurem. Esimene tulp *Dim1* koosneb kas 0-st või numbritest 1 ja 2 vastavalt sellele, mis neil oli tulbas *Best*. Me paneme sinna 1 või 2 vaid juhul, kui see rida kuulub parima rea gruppi (summeerides annavad kokku vähemalt 0,8). Teisel juhul paneme sinna 0. Seega ainult mittenullilised elemendid tuleks arvesse võtta telgede tõlgendamisel. Esimeses tulbas on vaid 7 mittenullilist elementi. Seega võib eeldada, et nendest 7-st võistlejast moodustunud grupp eristub teistest esimesel teljel.

Tabel 6. Reaprofiilide koosinuste ruudud

	Dim1	Dim2
EATON	0.0029	0.4417
HARDEE	0.0877	0.1459
SUAREZ	0.6610	0.0760

VANALPHE	0.2645	0.0590
WARNER	0.1609	0.0514
FREIMUTH	0.3435	0.0009
KASYANOV	0.1942	0.0454
SVIRIDOV	0.3505	0.3095
COERTZEN	0.4705	0.0857
BEHRENBR	0.1022	0.0570
SINTNICO	0.0608	0.7573
NEWDICK	0.0635	0.0868
BARROILH	0.0046	0.6339
GARCIA	0.0017	0.0008
MAYER	0.6963	0.0036
SHKURENE	0.0088	0.5583
MIKHAN	0.2102	0.1487
KARPOV	0.6172	0.1228
DEARAUJO	0.1512	0.0005
USHIRO	0.4435	0.0033
VOS	0.0268	0.0815
ERINS	0.1424	0.2159
ADDY	0.7662	0.1480
SZABO	0.0626	0.1249
DRAUDVIL	0.6137	0.2449
ARTIKOV	0.2131	0.0638

Tabel 6 eraldab kvaliteedi mõõtmed kahe telje vahel. Rea i jaoks annab see θ_{ik} koosinuse, mis on nurk reaprofiili vektori (alguspunktiks keskpunkt) ja selle k-ndale teljele moodustunud projektsiooni vahel:

$$\cos\theta_{ik}^2 = \frac{x_{ik}^2}{d^2(i)}\,.$$

Täheldatav on asjaolu, et kahe koosinuse ruutude summa, $\cos\theta_{i1}^2 + \cos\theta_{i2}^2$, annab kokku meile i-nda rea kahedimensonaalse esituse kvaliteedi. Võib veel märkida, et koosinuse ruutude summa üle kõikide telgede koordinaatide *k = 1,...,K* annab meile tulemuseks 1.

Järgnevas toome ära analoogse analüüsi veeruprofiilide jaoks.

Tabel 7. Veeruprofiilide koordinaadid

	Dim1	Dim2
a1	-0.0391	0.0030
a2	0.0070	-0.0034
a3	-0.0433	0.0116
a4	0.0163	-0.0019
a5	-0.0142	0.0066
a6	-0.0436	-0.0124
a7	-0.0314	0.0684
a8	0.0031	-0.0912

a9	0.0887	0.0297
a10	0.0835	0.0007

Tabel 8. Veeruprofiilide statistikud

	Quality	Mass	Inertia
a1	0.4617	0.1094	0.0621
a2	0.0174	0.1075	0.0633
a3	0.4151	0.0941	0.0776
a4	0.1012	0.0983	0.0443
a5	0.1263	0.1055	0.0347
a6	0.5061	0.1130	0.0781
a7	0.6937	0.0943	0.1310
a8	0.8463	0.1029	0.1722
a9	0.7442	0.0912	0.1825
a10	0.6432	0.0837	0.1543

Tabel 9. Veeruprofiilide panus telgedele

	Dim1	Dim2
a1	0.0835	0.0007
a2	0.0026	0.0009
a3	0.0880	0.0089
a4	0.0129	0.0003
a5	0.0106	0.0032
a6	0.1071	0.0123
a7	0.0462	0.3120
a8	0.0005	0.6048
a9	0.3578	0.0568
a10	0.2908	0.0000

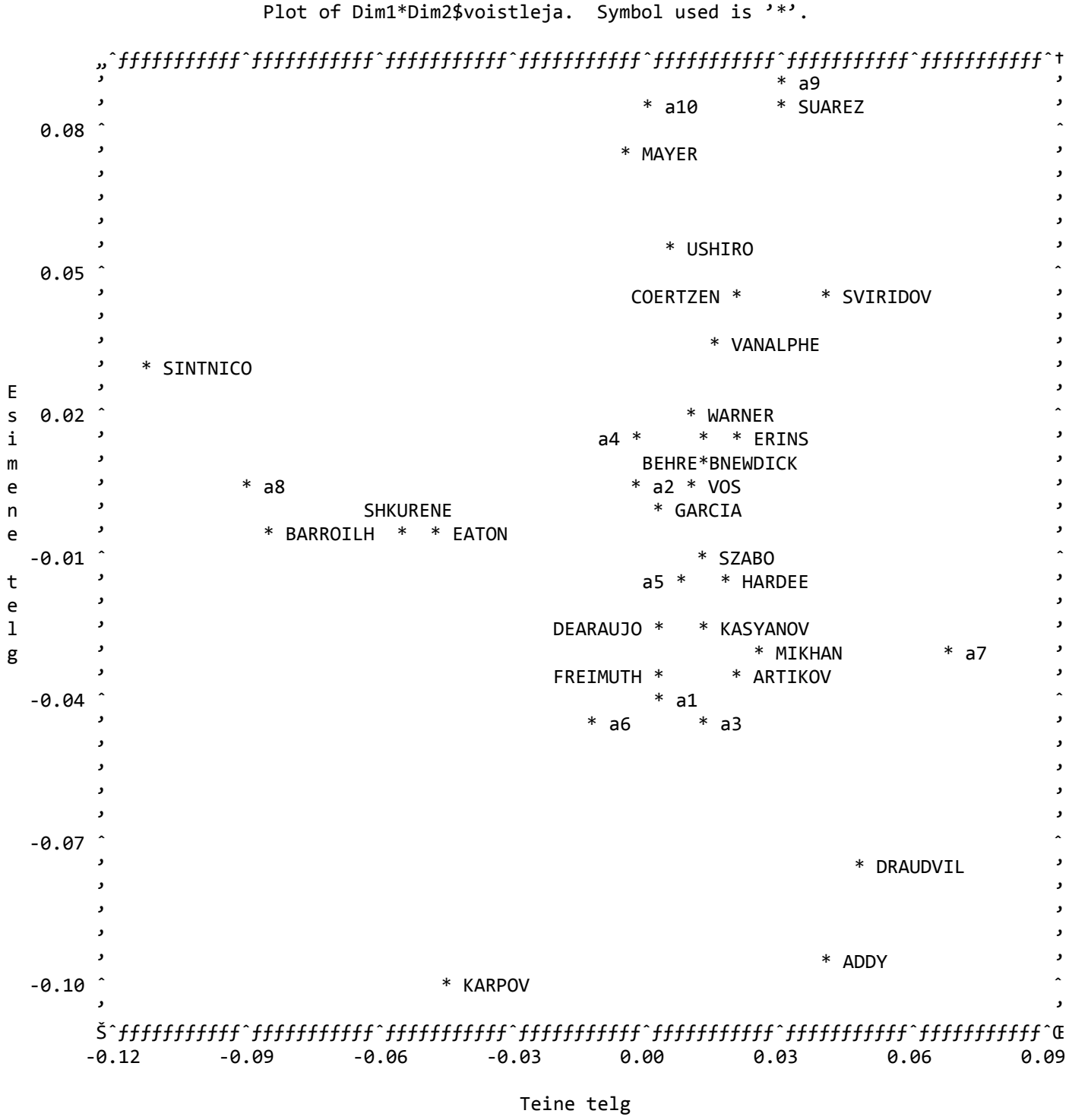
Tabel 10. Millised alad iseloomustavad kõige paremini telgi

	Dim1	Dim2	Best
a1	0	0	1
a2	0	0	1
a3	1	0	1
a4	0	0	1
a5	0	0	1
a6	1	0	1
a7	0	2	2
a8	0	2	2
a9	1	0	1
a10	1	0	1

Tabel 11. Veeruprofiilide koosinuse ruudud

	Dim1	Dim2
a1	0.4589	0.0028
a2	0.0140	0.0034
a3	0.3874	0.0277
a4	0.0998	0.0014
a5	0.1039	0.0224
a6	0.4681	0.0380
a7	0.1204	0.5732
a8	0.0009	0.8453
a9	0.6693	0.0749
a10	0.6432	0.0001

Viimased viis tabelit sisaldavad tulemusi veeruprofiilide (alade) analüüsi kohta. Neid tabeleid saab kirjeldada täpselt samal viisil nagu reaprofiilide korral. Tabelist 8 võib lugeja märgata, et a8 (teivashüpe) kvaliteet on üle 0,8, mis tähendab, et teivashüpe ei vaja rohkem telgi punkti kirjeldamiseks. Samas võib märgata, et a2 (kaugushüpe) kvaliteet on üsna nullilähedane, ehk see punkt vajaks parema kirjelduse saamiseks kolmandat või rohkemat telge. Tabelist 9 võib lugeja märgata, et a8 (teivashüpe) on väga hästi kirjeldatud teise peatelje poolt ning a9 ja a10 (vastavalt odavise ja 1500m jooks) on väga hästi kirjeldatud esimese peatelje poolt.



Joonis 2. Võistlejate ja alade kombineeritud graafik

Lõpuks toome ära graafilise esituse, kus on nii võistlejad, kui ka alad. Antud graafik on saadud vastavalt koordinaatidele, mis on eelnevates tabelites välja toodud.

On kergesti märgatav, et andmestik on heterogeenne, sisaldades endas erinevaid klastreid. Nagu jooniselt paistab, siis moodustavad enamus võistlejad ja ka spordialad keskele ühe suure klastri. Graafiku vasakus küljes on võistlejad Sintnicolaas, Barroilhet, Shkurenev ja Eaton ning ka 8. võistlusala, milleks on teivashüpe. Nimetatud 4 sportlast ongi head teivashüppajad. Graafiku ülemise osa moodustavad võistlejad Suarez, Mayer, Ushiro, Coertzen, Svirodov ja Van Alphen ning ka 9. ja 10. ala (vastavalt odavise ja 1500m jooks), mis on nende sportlaste tugevad küljed. Need alad pole kõige paremad aga võistlejatele Karpov, Addy ja Draudvila, kes paiknevad joonise allosas. Eriliseks võib pidada ka graafiku paremal pool asetsevat punkti a7, milleks on kettaheide. Silma paistab ka, et keskmisse klastrisse on jäänud kõik esimese päeva alad. Asjaolu, et esimese päeva alad paiknevad suhteliselt lähestikku tähendab seda, et need alad korreleeruvad tugevasti – kes sooritab hästi ühe ala, sooritab reeglina hästi ka teised alad.

Moodustatud kaks telge on head tõlgendamaks erinevusi klastrite vahel, küll aga on keerukam tõlgendada klastrisiseseid erinevusi.

2.2. Rakendus andmete dubleerimisega

Ka selles analüüsi osas kasutame sama näidet. Küll aga läheneme siin andmestikule veidi teise nurga alt. Enamasti mõõdavad „positiivselt orienteeritud“ skaalad seda, kui hästi keegi midagi sooritas. Näiteks meie näite puhul seda, kui palju punkte mingi sportlane oma soorituse eest teenis. Samas aga saaks ekvivalentselt kasutada ka „negatiivselt orienteeritud“ skaalat, mis näitab kui palju jättis sportlane mingil alal punkte võtmata. Pööratud skaala puhul lahutan alade tulemused igal sportlasel maha vastavalt iga ala kahekordsest keskmisest. Seega tekivad uued tunnused: a_1=1758-a1, a_2=1727-a2, a_3=1512-a3, ..., a_10=1345-a10, kokku_=16065-kokku.

Sellist andmestiku täiendamist „peegeldatud“ väärtustega nimetakse andmestiku dubleerimiseks. Meie dubleeritud andmestikus on meil seega 26 rida(võistlejad) ja 10+10=20 veergu.

Selles näites on lisaks kasutatud veel ühte „trikki“, et anda paremini aimu andmestikust. Selleks on täiendavate ridade ja veergude meetod. Idee seisneb selles, et kui profiilide kaudu on kaks olulisemat telge kindlaks tehtud, siis on võimalik graafiliselt kujutada veel mõningaid punkte, samal ajal neid oma arvutustesse lisamata. Meie näites ei võta me kogusummat telgede leidmisel arvesse, aga oleks huvitav lisada see graafikule, et seda saaks seostada teiste profiilidega ja telgedega.

Tabel 12. Inertsid ja hii-ruut statistiku lahutamine dubleeritud andmestiku korral

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	7 14 21 28 35
0.05379	0.00289	1208.31	36.28	36.28	************************
0.04006	0.00160	670.33	20.12	56.40	**********
0.03628	0.00132	549.81	16.51	72.91	**********
0.02762	0.00076	318.75	9.57	82.48	*******
0.02170	0.00047	196.64	5.90	88.38	****
0.01770	0.00031	130.82	3.93	92.31	***
0.01542	0.00024	99.33	2.98	95.29	**
0.01523	0.00023	96.87	2.91	98.20	**
0.00966	0.00009	39.00	1.17	99.37	*
0.00710	0.00005	21.04	0.63	100.00	
Total	0.00797	3330.90	100.00		

Degrees of Freedom = 475

Nagu näha, siis kaks esimest telge kirjeldavad andmestikku 56,4%. See on peaaegu sama hea nagu näite esimeses pooles.

Tabel 13. Reaprofiilide statistikud dubleeritud andmestikus

	Quality	Mass	Inertia
EATON	0.9582	0.0385	0.0834
HARDEE	0.4220	0.0385	0.0424
SUAREZ	0.8015	0.0385	0.0805
VANALPHE	0.4960	0.0385	0.0350
WARNER	0.6259	0.0385	0.0253
FREIMUTH	0.3141	0.0385	0.0248
KASYANOV	0.3306	0.0385	0.0215
SVIRIDOV	0.5851	0.0385	0.0300
COERTZEN	0.5613	0.0385	0.0240
BEHRENBR	0.2162	0.0385	0.0123
SINTNICO	0.2595	0.0385	0.0794
NEWDICK	0.1781	0.0385	0.0087
BARROILH	0.1032	0.0385	0.0560
GARCIA	0.0590	0.0385	0.0129
MAYER	0.4886	0.0385	0.0376
SHKURENE	0.2011	0.0385	0.0253
MIKHAN	0.1990	0.0385	0.0187
KARPOV	0.6657	0.0385	0.0773
DEARAUJO	0.2308	0.0385	0.0220
USHIRO	0.5471	0.0385	0.0363
VOS	0.2384	0.0385	0.0105
ERINS	0.5342	0.0385	0.0198
ADDY	0.7716	0.0385	0.0637
SZABO	0.7892	0.0385	0.0210
DRAUDVIL	0.7379	0.0385	0.0545
ARTIKOV	0.8914	0.0385	0.0769

Kvaliteedi näitaja on väga suur (rohkem kui 0,8) kolmel sportlasel, kelleks on Eaton, Suarez ja Artikov. Samas sportlaste Barrouilhet, Garcia ja Newdick kvaliteet on lausa alla 0,2. Seega võivad nemad vajada kolmandat või ka rohkemat telge, et saada paremini graafikul kirjeldatud. Suurimad inertsid on sportlastel Eaton, Suarez ja Sintnicolaas. See on seetõttu, et nende profiilid erinevad keskmisest kõige rohkem. Samas see on ka loogiline, kuna Eaton on meie võistluse võitja. Suarez esines väga hästi odaviskes ja Sintnicolaas teivashüppes, mis eristas neid teistest.

Tabel 14. Veeruprofiilide statistikud dubleeritud andmestikus

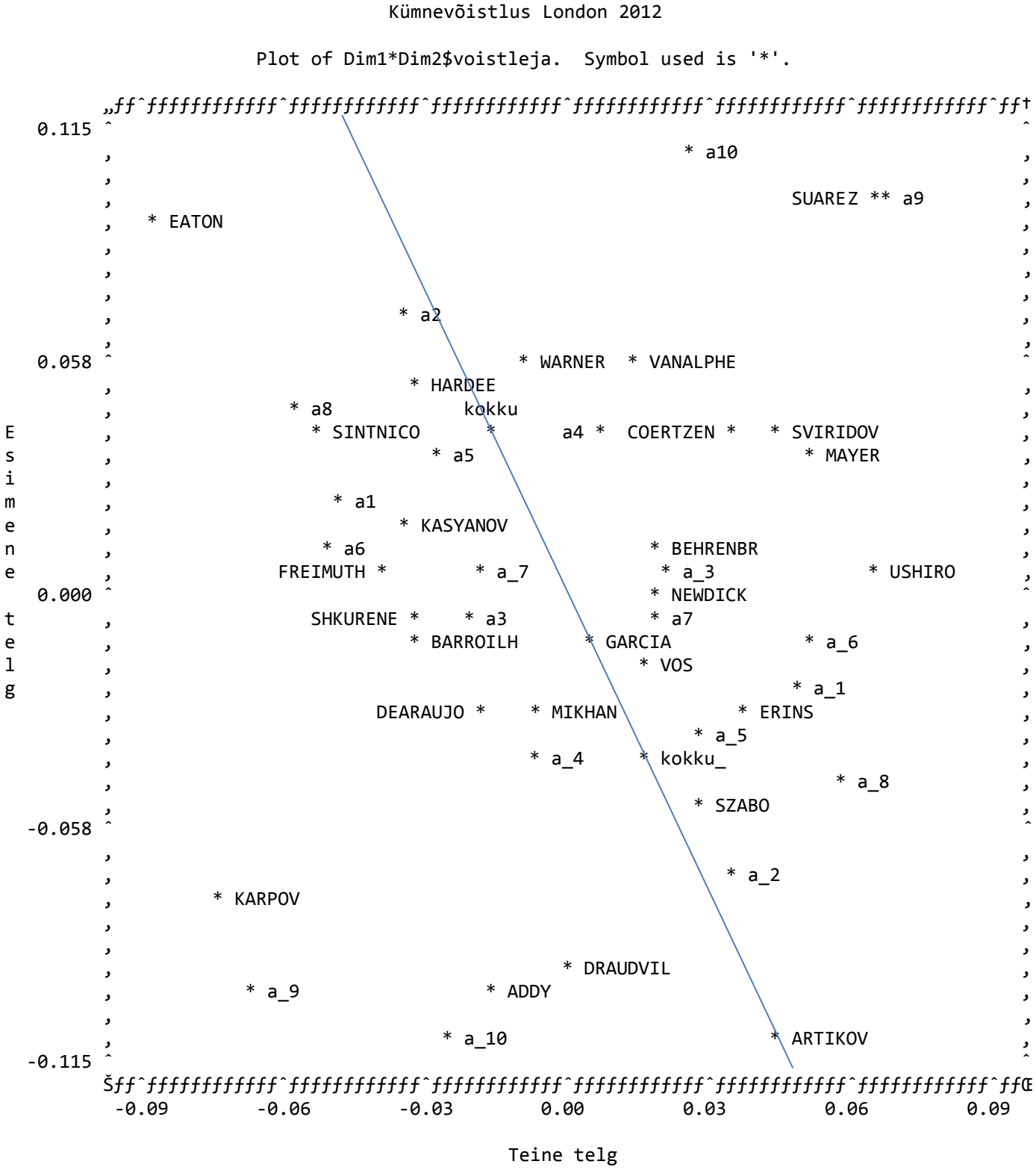
	Quality	Mass	Inertia
a1	0.5532	0.0547	0.0344
a2	0.7053	0.0538	0.0574
a3	0.1079	0.0471	0.0237
a4	0.4370	0.0491	0.0247
a5	0.5231	0.0528	0.0258
a6	0.5197	0.0565	0.0373
a7	0.0436	0.0471	0.0441
a8	0.4484	0.0514	0.0780
a9	0.8324	0.0456	0.0919
a10	0.7607	0.0419	0.0826
a_1	0.5532	0.0547	0.0344
a_2	0.7053	0.0538	0.0574
a_3	0.1079	0.0471	0.0238
a_4	0.4370	0.0491	0.0247
a_5	0.5231	0.0527	0.0258
a_6	0.5197	0.0565	0.0374
a_7	0.0436	0.0472	0.0441
a_8	0.4484	0.0515	0.0780
a_9	0.8324	0.0456	0.0919
a_10	0.7607	0.0419	0.0826

Nagu tabelist näha, on kolme ala a2, a_2 (kaugushüpe), a9, a_9 (odavise), a10 ja a_10 (1500m jooks) veeruprofiilide kvaliteet kõige parem (üle 0,7). Samal ajal aga a7 ja a_7 (kettaheide) ning a3 ja a_3 (kuulitõuge) kvaliteet on väga madal ja vajaks kolmandat või rohkemat telge, et paremini kirjeldatud saada. Suurimad inertsid on veeruprofiilidel a9 ja a10, milleks on odavise ja 1500m jooks.

Tabel 15. Veeruprofiilide koosinuse ruudud dubleeritud andmestikus

	Dim1	Dim2
a1	0.0833	0.4699
a2	0.5694	0.1358
a3	0.0067	0.1012

a4	0.4284	0.0087
a5	0.3170	0.2061
a6	0.0349	0.4847
a7	0.0012	0.0425
a8	0.1824	0.2660
a9	0.5624	0.2700
a10	0.7220	0.0386
a_1	0.0833	0.4699
a_2	0.5694	0.1358
a_3	0.0067	0.1012
a_4	0.4284	0.0087
a_5	0.3170	0.2061
a_6	0.0349	0.4847
a_7	0.0012	0.0425
a_8	0.1824	0.2660
a_9	0.5624	0.2700
a_10	0.7220	0.0386



NOTE: 1 obs had missing values.

Joonis 3. Dubleeritud andmestiku võistlejate ja alade kombineeritud graafik

Joonisel 3 on kujutatud dubleeritud andmestiku kõik võistlejad ja alad. Nagu näha, siis on esimene peatelg peamiselt defineeritud 10. ala kaudu, milleks on 1500m jooks. Nende korrelatsiooni ruutu 0,72 saab vaadata tabelist 15. Teine telg pole märgatavalt milligagi korreleeritud (kõik alla 0,5).

Et esimene telg on kõige rohkem korreleeritud 1500m jooksuga, siis meie näites, mida ülevamal pool asub võistleja, seda parem pikamaajooksja ta on. Et võistlejad Suarez ja Eaton on selgelt kõige ülemised, siis võib pidada neid väga heaks 1500m jooksjaks. Allpool graafiku osas asuvad võistlejad, kes on esirinnas aladel saamata jäänud punktide osas. Sinna kuuluvad Artikov, Addy, Draudvila ja Karpov ning alad a_10, a_9 ja a_2. Nende võistlejate kohta võib väita, et neil ebaõnnestusid väga paljud alad vastavalt: 1500m jooks, odavise ja kaugushüpe. Ka siin graafikul on näha, et võistlejat Suarez võib pidada väga heaks odaviskajaks ja 1500m jooksjaks.

Tabel 16. Lisa veeruprofiilide kvaliteet

kokku	0.8953
kokku_	0.8953

Meie näites defineerivad punktid kokku ja kokku_ uue telje, mis asetseb kahe esialgse telje vahel, olles samal ajal väga lähedane peatasandiga, sest kvaliteet on peaaegu 0,9 (Tabel 16). Seega, võistlejate projektsioonid täiendavale teljele tekitaks võistlejate järjestuses korrektuure.

Võib küsida, miks on korrespondentsanalüüsiga koostatud järjestus erinev esialgsest? See on aga seetõttu, et võistlejad, kes mingil alal saavutavad väga hea tulemuse, samal ajal kui teised olid tol alal märgatavalt nõrgemad, saab võistleja rohkem punkte, kui tema tegelik punktisaak. Seda seetõttu, et tema õnnestus alal, kus teised ebaõnnestusid. Selline seaduspära kehtib ka vastupidisel juhul. Näiteks, kui keegi ebaõnnestub alal, kus teised õnnestuvad, siis saab ta vähem punkte, kui tema tegelik punktisaak olema peaks.

Meie näites tõuseks sellist korrespondentsanalüüsi kasutades võistleja Sintnicolaas paremusjärjestuses ettepoole, kuna tema teivashüpe oli teistest märgatavalt parem.

Kokkuvõte

Käesoleva bakalaureusetöö eesmärgiks oli anda ülevaade korrespondentsanalüüsi matemaatilisest meetodist. Töö rakenduslikus osas vaatlesime lähemalt 2012. aasta Londoni Olümpiamängude kümnevõistluse tulemusi ning proovisime tuvastada seoseid võistlejate ja alade vahel. Vaatluse tulemusena tuvastasime erinevaid klastreid. Tekkinud klastrite põhjal saame väita, et võistlejad Sintnicolaas, Barrouilhet, Shkurenev ja Eaton on väga head teivashüppajad. Võistlejad Suarez, Mayer, Ushiro, Coertzen, Svidorov ja Van Alphen on väga head odaviskad ja 1500m jooksjad. Viimati nimetatud alad pole kõige paremad võistlejatele Karpov, Addy ja Draudvila. Lisaks paistis silma, et kõige suuremasse klastrisse jäid kõik esimese päeva alad.

Rakendasime samade vaatlusandmetega ka andmete dubleerimist. Dubleeritud andmestiku graafikul tuvastasime, et võistlejad Artikov, Addy, Draudvila ja Karpov ei ole kõige paremad 1500m jooksjad, odaviskajad ja kaugushüppajad. Võistlejat Suarez võib pidada väga heaks odaviskajaks ja 1500m jooksjaks. Lisaks on seal graafikult täiendav telg, mille peale projektsioon teeks võitlejate järjestuses korrektuure. Nimelt tõuseks võistleja Sintnicolaas pingereas ettepoole.

Seega kokkuvõtteks võib öelda, et töö täitis oma eesmärki. Suutsime leida seoseid võistlejate ja alade vahel. Andmete dubleerimisega saime teada uusi seoseid.

Correspondence analysis and data doubling

Bachelor thesis

Gertis Aru

Summary

The aim of this bachelor thesis is to introduce the main aspects of the correspondence analysis and to show the mathematical theory that stands behind the method. At the end some examples are shown.

Correspondence analysis is a multivariate statistical technique. It is similar to principal component analysis and can be used to reveal the main dimensions in the data. Correspondence analysis can be applied not only for two-way contingency tables but to analyze extremely large variety of data that can be brought into the form of two-way table of non-negative numbers.

We have I rows and J columns and thus the data can be presented in the form of IxJ contingency table. For each row and column, we have to find their profile. Every profile has its own mass. All row profiles and column profiles construct a cloud. Centroid is the weighted average of all elements in the cloud.

Our main aim is to identify a low-dimensional subspace which comes closest to all points in the cloud. The important concept in correspondence analysis is inertia. Inertia is the measure of how much the profiles are spread around the centroid. So that a specific low-dimensional sub-space which comes close to all the points of the cloud is determined by principal axes of inertia. For most cases, we want a two-dimensional display of row and column profiles. Therefore, we need only two first dimensions to depict the data.

Mathematical solution to the problem is achieved by using singular value decomposition. By using that, we identify the largest eigenvalues which describe our principal axes. Every row and column profile have principal co-ordinates which are described by the principal axes. So, by using the principal co-ordinates and principal axes, we can display the data on a two-dimensional graph.

In the second part we analyse 2012 London Olympics decathlon results. It is seen that there are different clusters. We can say that athletes Sintnicolaas, Barrouilhet, Shkurenev and Eaton are great at pole vault. Athletes Suarez, Mayer, Ushiro, Coertzen, Svidorov and Van Alphen are good at javelin throw and 1500m running. Whereas athletes Karpov, Addy and Draudvila are not so good at those events. It is also seen that all the first day events are in a big cluster.

Also we use doubling of the data. On the doubled data graph it is seen that athletes Artikov, Addy, Draudvila and Karpov are not good at 1500m running, javelin throw nor at long jump. Athlete Suarez can be considered to be great at javelin throwing and 1500m running. We also use a supplementary axis and the projections of the athletes' points on that supplementary axis create a different ordering. For example athlete Sintnicolaas would obtain a better rank.

Lisad

Lisa 1. Programm kümnevõistluse tulemuste analüüsiks

```
title 'Kümnevõistlus London 2012';
data kymnevoistlus;
input voistleja$ a1-a10 kokku;
cards;
EATON 1011 1068 769 850 963 1032 716 972 767 721 8869
HARDEE 994 942 807 794 904 1035 834 849 838 674 8671
SUAREZ 801 940 759 906 859 917 782 819 996 744 8523
VANALPHEN 850 970 819 850 853 863 835 849 763 795 8447
WARNER 980 945 712 850 899 926 785 819 780 746 8442
FREIMUTH 940 864 782 714 906 989 852 880 698 695 8320
KASYANOV 961 947 756 794 888 963 802 790 661 721 8283
SVIRIDOV 910 922 754 794 866 799 817 790 865 702 8219
COERTZEN 841 854 715 850 882 955 738 760 810 768 8173
BEHRENBRUCH 847 850 831 767 813 932 761 819 810 696 8126
SINTNICOLAAS 894 903 739 740 868 920 509 1004 720 737 8034
NEWDICK 838 900 795 767 804 847 791 819 735 692 7988
BARROILHET 821 767 758 850 766 959 690 1035 697 629 7972
GARCIA 906 755 758 794 873 944 711 790  736 689 7956
MAYER 791 854 731 850 873  780 689 819 774 791 7952
SHKURENEV 858 874 661 822 823 925 736 941 645 663 7948
MIKHAN 919 799 774 740 889 955 755 731 673 693 7928
KARPOV 881 864 880 794 822 924 765 941 588 467 7926
DEARAUJO 929 852 699 740 897 875 762 790 612 693 7849
USHIRO 791 781 703 794 779 794 801 880 834 685 7842
VOS 865 878 714 767 832 897 711 760 762 619 7805
ERINS 863 809 695 740 786 823 769 760 698 706 7649
ADDY 885 790 788 740 878 945 779 673 594 514 7586
SZABO 827 804 724 714 777 859 770 790 720 596 7581
DRAUDVILA 872 842 800 767 809 865 796 673 591 542 7557
ARTIKOV 780 677 735 740 729 881 737 731 687 506 7203
;
proc corresp outc=coor;
var a1--a10;
id voistleja;
run;
proc plot;
plot dim1 * dim2 = '*' $ voistleja/
     box haxis= -0.11 to 0.09 by 0.04 vaxis=-0.11 to 0.09 by 0.04;
label dim1='Esimene telg’
      dim2='Teine telg';
run;
quit;
```

Lisa 2. Programm dubleeritud andmestiku analüüsiks

```
title 'Kümnevõistlus London 2012 (osa2)';
data kymnevoistlus;
input voistleja$ a1-a10 kokku;
a_1=1758-a1;
a_2=1727-a2;
a_3=1512-a3;
a_4=1579-a4;
a_5=1695-a5;
a_6=1815-a6;
a_7=1515-a7;
a_8=1653-a8;
a_9=1466-a9;
a_10=1345-a10;
kokku_=16065-kokku;
cards;
EATON 1011 1068 769 850 963 1032 716 972 767 721 8869
HARDEE 994 942 807 794 904 1035 834 849 838 674 8671
SUAREZ 801 940 759 906 859 917 782 819 996 744 8523
VANALPHEN 850 970 819 850 853 863 835 849 763 795 8447
WARNER 980 945 712 850 899 926 785 819 780 746 8442
FREIMUTH 940 864 782 714 906 989 852 880 698 695 8320
KASYANOV 961 947 756 794 888 963 802 790 661 721 8283
SVIRIDOV 910 922 754 794 866 799 817 790 865 702 8219
COERTZEN 841 854 715 850 882 955 738 760 810 768 8173
BEHRENBRUCH 847 850 831 767 813 932 761 819 810 696 8126
SINTNICOLAAS 894 903 739 740 868 920 509 1004 720 737 8034
NEWDICK 838 900 795 767 804 847 791 819 735 692 7988
BARROILHET 821 767 758 850 766 959 690 1035 697 629 7972
GARCIA 906 755 758 794 873 944 711 790  736 689 7956
MAYER 791 854 731 850 873  780 689 819 774 791 7952
SHKURENEV 858 874 661 822 823 925 736 941 645 663 7948
MIKHAN 919 799 774 740 889 955 755 731 673 693 7928
KARPOV 881 864 880 794 822 924 765 941 588 467 7926
DEARAUJO 929 852 699 740 897 875 762 790 612 693 7849
USHIRO 791 781 703 794 779 794 801 880 834 685 7842
VOS 865 878 714 767 832 897 711 760 762 619 7805
ERINS 863 809 695 740 786 823 769 760 698 706 7649
ADDY 885 790 788 740 878 945 779 673 594 514 7586
SZABO 827 804 724 714 777 859 770 790 720 596 7581
DRAUDVILA 872 842 800 767 809 865 796 673 591 542 7557
ARTIKOV 780 677 735 740 729 881 737 731 687 506 7203
;
proc corresp outc=coor;
var a1--kokku_;
supplementary kokku kokku_;
id voistleja;
run;
proc plot;
plot dim1 * dim2 = '*' $ voistleja/
     box haxis= -0.09 to 0.09 by 0.03 vaxis=-0.115 to 0.115 by 0.0575;
label dim1='Esimene telg'
      dim2='Teine telg';
run;
quit;
```

Viited

1. Pärna, K. (1993). Correspondence Analysis: An Introduction and Some Examples. Stockholm.
2. Greenacre, M.J. (1984). Theory and Application of Correspondence Analysis. Academic Press, London.
3. Greenacre, M.J. (1993). Correspondence Analysis in Practice. Academic Press, London.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina Gertis Aru
(*autori nimi*)

(sünnikuupäev: 22.11.1990)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
Korrespondentsanalüüs ja andmete dubleerimine,

(*lõputöö pealkiri*)

mille juhendaja on prof. Kalev Pärna,
(*juhendaja nimi*)

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, **06.05.2013**