TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
MATEMAATILISE STATISTIKA INSTITUUT

Külli Koov

TARTU ÜLIKOOLI TÖÖTAJATE ANDMESTIKU KORRESPONDENTSANALÜÜS

Bakalaureusetöö

Juhendaja: Tõnu Kollo

TARTU 2004

Sisukord

Sissejuhatus

Korrespondentsanalüüsi kui visualiseerimismeetodi eesmärk on leida graafiline esitus, mille tõlgendamine võimaldaks kindlaks teha, millised on tunnustevahelised seosed ja neid seoseid iseloomustavad tasemed. Korrespondentsanalüüs võimaldab illustreerida suuri sagedustabeleid kahemõõtmelise (kolmemõõtmelise) graafiku abil. Kuna suuri andmestikke on tülikas hajuvusdiagrammina esitada, siis on võimalik seejuures saada palju ülevaatlikum ettekujutus seoste olemusest.

Korrespondentsanalüüs töötati välja 1960-ndatel ja 1970-ndatel aastatel Prantsusmaal lingvistika probleemide uurimiseks. Hiljem muutus see populaarseks ka ingliskeelsetes maades ja leidis laialdast kasutamist muudes valdkondades.

Käesoleva bakalaureusetöö esimene osa sisaldab korrespondentsanalüüsi ülevaadet, põhimõistete tutvustust ning korrespondentsanalüüsi kasutamise üldist kirjeldust. Teises töö osas on korrespondentsanalüüsi rakendatud Tartu Ülikooli töötajaskonna andmestikule ja interpreteeritud saadud tulemusi. Antud töö on jätkuks semestritööle Koov (2004). Tööl on kaks lisa, kus on toodud korrespondentsanalüüsi tulemused ja programmid, millega korrespondentsanalüüsi tulemused saadi.

Korrespondentsanalüüsi läbiviimiseks kasutame andmetöötlusprogrammi SAS.

1. Korrespondentsanalüüs

1.1 Üldiseloomustus

Käesoleva töö korrespondentsanalüüsi üldiseloomustus ja meetodi kirjeldus on kirjutatud K. Pärna uurimuse (Pärna, 1993) alusel.

Korrespondentsanalüüs on statistikameetod analüüsimaks kahemõõtmelisi andmetabeleid, mille tulemusena esitatakse tabeli read ja veerud punktidena madalamamõõtmelises vektorruumis.

Hästi on teada, et 'klassikalised' statistilised meetodid on välja töötatud peamiselt analüüsimaks arvandmeid. Tänapäeval kasutatakse statistikat väga erinevates valdkondades ja tihti püüavad uurijad kirjeldada uuritavaid probleeme mittearvuliste tunnustega. On teada, et tegelikult sotsiaalteadustes, maamõõtmis-uuringutes, turu-uuringutes ja mujal kasutatakse peaaegu alati ka mittearvulisi andmeid. Mõningad näited sellistest tunnustest oleksid: ametinimetus, kodumaa, auto tüüp, piirkond, sugu, usuline kuuluvus, rass, nõustumine/mittenõustumine antud väitega jne. Ühtki nendest tunnustest ei saa otseselt arvuliselt mõõta. Neile väärtustele saame vaid seada vastavusse konkreetsed koodid (numbrid) ja kasutada neid koode sobival viisil analüüsimaks andmeid. Andmete puhul eristame kahte mõõtmistaset. Kui juhuslike suuruste väärtuste vahel ei ole mingit loomulikku järjestust, siis võib koodid valida suvaliselt. Sellisel juhul ütleme, et tunnused on mõõdetud nominaalskaalal. Aga kui tunnuste väärtustel eksisteerib järjestus mingi sisulise tähendusega, siis koodid peaksid järgima sama järjestust. Sellisel juhul ütleme, et tunnused on mõõdetud järjestusskaalal. Nõustumise/mittenõustumise tunnused ja haridustase on klassikalised näited sellistest tunnustest.

Klassifitseeritud andmed on tüüpiliselt esitatud sagedustabelina. Sagedustabelite analüüs vajab erilisi statistilisi meetodeid, mis erinevad harilikest korrelatsiooni- ja/või regressioonipõhistest tehnikatest analüüsimaks kvalitatiivseid andmeid. Traditsioonilise kahemõõtmelise sagedustabeli analüüs sisaldab lahtrite loendit,

protsente, χ^2-statistikut testimaks sõltumatust tunnuste vahel ja mitmeid muid erinevaid seosemõõte. Mitmemõõtmeline sagedustabel vajab keerukamat lähenemist. On olemas üldine viis käsitleda mõlemat, nii kahemõõtmelist, kui ka mitmemõõtmelist tabelit – loglineaarne mudel. Loglineaarseid mudeleid on kasutatud ulatuslikult alates 1970. aastatest, suuresti tänu L. Goodmani, S. Habermanni ja teiste töö tulemusena. Loglineaarne mudel on üldine termin mitme erineva mudeli kohta: mudel log-sageduse loendamiseks k-mõõtmelises (k=2 või suurem) sagedustabelis; logitmudel juhuks, kui üks tunnus on sõltuv ja teine on uuritav; ja palju teisi erinevaid mudeleid.

Loglineaarsetel mudelitel on palju häid omadusi. Neid saab rakendada (vähemalt teoorias) suvaliste mõõtmetega tabelitele. Nad annavad uurijale sobivusstatistiku, mille abil saab leida sobiva mudeli ja ka parameetrite hinnangud ning nende standardvead.

Aga siiski võib välja tuua ka mõningaid puudusi loglineaarse mudeli puhul. Üheks probleemiks on null elementide loendis. Samuti eeldab χ^2-statistik väikest valimimahtu. Kui vastupidiselt, valimimaht on väga suur, siis on raske leida lihtsat mudelit. Probleem tekib ka siis, kui muutujate arv on suur või nendel muutujatel on palju väärtusi. See tähendab, et peab kasutama väga palju parameetreid, mis võib viia raskusteni interpreteerimisel ja parameetrite hindamisel.

Korrespondentsanalüüs on tehnika, mida võib kasutada loglineaarmudeli täiendina või selle asemel. Kõige lihtsamal viisil saab seda rakendada lähtudes kahemõõtmelisest sagedustabelist, saades tulemuseks arvulised väärtused nii rea- kui ka veeruklassides. Need väärtused valitakse selliselt, et oleks võimalik kõige paremini kirjeldada kahe tunnuse vahelist seost. Tavaliselt rea- ja veeruklassid esitatakse kahemõõtmelisel joonisel väärtuste paaridena. Selline esitus annab uurijale visuaalse kujutuse andmetest ja sealt on näha erinevusi ja sarnasusi rea- ja veeruklassides. Meetodil on mitmeid sarnasusi peakomponentide analüüsiga ja seda saab kasutada esile toomaks põhilisi 'dimensioone' andmetes. Siinkohal mainime ka, et korrespondentsanalüüsi saab rakendada mitte ainult kahemõõtmelisele sagedustabelile, aga ka analüüsimaks äärmiselt suuri mitmekesiseid andmeid, mille saab viia kahemõõtmelise tabeli kujule.

Eelduseks on, et need andmed ei sisalda mittenegatiivseid suurusi. Uurijad kasutavad mitmeid erinevaid arvutiprogramme tegemaks korrespondentsanalüüsi. Näiteks statistikapaketid SAS, SPSS, BMDP võimaldavad läbi viia korrespondentsanalüüsi.

1.2 Korrespondentsanalüüsi meetodi kirjeldus

1.2.1 Tähistused ja mõisted

Olgu meil n vaatlust (inimesed, objektid) klassifitseeritud kahe juhusliku muutuja põhjal klassi A ja B. Olgu muutujal A kokku I klassi $A_1, A_2, ..., A_I$ ja muutujal B olgu J klassi $B_1, B_2, \ldots, B_J$. Olgu $n_{i\cdot}$ vaatluste arv, mille väärtused on A_i ja sarnaselt $n_{\cdot j}$ on vaatluste arv, mille väärtusteks on B_j. Vaatluste arv, millel on ühesugused väärtused A_i ja B_j märgistatakse tähisega n_{ij}. Seega saab andmed esitada $I \times J$ sagedustabeli $N = (n_{ij})$ kujul.

Sagedustabel N näeb välja järgnevalt

	B_1	B_2	$\cdots$	B_J	$\sum$
A_1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I\cdot}$
$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot J}$	n

Siis tabeli N kõigi elementide summa on

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

ja ridade ning veergude kogusummad on vastavalt

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

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

Korrespondentsanalüüsis on meie peamiseks huviobjektiks sagedustabeli N tinglikud jaotused. Tähistame järgnevad suhtelised sagedused (tõenäosuste hinnangud) tabelis N järgnevalt:

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

$$f_i = \frac{n_{i\cdot}}{n},$$

$$f_j = \frac{n_{\cdot j}}{n}.$$

Korrespondentsanalüüsis on tähtsal kohal mõiste profiil. Toome siinkohal sisse mõisted reaprofiil ja veeruprofiil. Vektor $f_B^i = (f_1^i, f_2^i, \ldots, f_J^i)^T$ esitab veerumuutuja B tingliku jaotuse eeldusel, et reamuutuja $A = A_i$. Me nimetame vektorit f_B^i *reaprofiiliks* (rea i jaoks). Reaprofiili element avaldub kujul

$$f_j^i = \frac{n_{ij}}{n_{i\cdot}} = \frac{f_{ij}}{f_i}.$$

Sarnaselt esitab vektor $f_A^j = (f_1^j, f_2^j, \ldots, f_I^j)^T$ reamuutuja A tinglikku jaotust eeldusel, et veerumuutuja $B = B_j$. Sellist tinglikku jaotust nimetatakse *veeruprofiiliks* (veeru j jaoks). Ja veeruprofiili element avaldub kujul

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

Terminil 'profiil' on see eelis, et teda saab rakendada igale mittenegatiivsete väärtustega ristkülikukujulisele tabelile N (ei pea olema sagedustabel), mis on üsna tüüpiline korrespondentsanalüüsis.

Näeme, et reaprofiilid f_B^i $(i = 1,2,\ldots,I)$ esitavad I punkti J-dimensionaalses eukleidilises ruumis R^J. Neid punkte nimetame I punkti *pilveks* ruumis R^J. Võtame kasutusele tähistuse $N_B(A)$, mis on reaprofiili pilv:

$$N_B(A) = \left\{ f_B^i \middle| i = 1,\ldots,I \right\}.$$

Võtame kasutusele veel ühe arvulise karakteristiku. Igale punkti f_B^i jaoks pilves $N_B(A)$ defineerime tema *massi*, kui rea i marginaalse tõenäosuse f_i. Seega on pilv

kaalutud punktide konfiguratsioon. See tähendab, et masside summa $\sum_{i=1}^{I} f_i = 1$. Selle abil defineeritakse pilve *tsentroid* kui tema masskese ehk pilve kõikide elementide kaalutud keskmine. Tähistame pilve $N_B(A)$ tsentroidi tähisega f_B. Seega, vastavalt definitsioonile on pilve $N_B(A)$ tsentroid:

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

mis on võrdne muutuja *B* marginaaljaotusega:

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

Vaatame nüüd dimensiooniga seonduvaid probleeme. Kõik pilve $N_B(A)$ punktid on elemendid *J*-mõõtmelises ruumist. Aga kuna vaadeldavate vektorite koordinaatide summa on üks, siis näeme, et profiilid tegelikult asetsevad alamruumis dimensiooniga (*J-1*). Teisest küljest sõltub alamruumi dimensioon punktide arvust *I*. Teame, et läbi kahe punkti saab panna sirge, kolm punkti asetsevad kahemõõtmelisel tasapinnal jne. Seega on selge, et pilved vajavad kirjeldamiseks mitte rohkem kui $\min\{J-1, I-1\}$-mõõtmelist ruumi. Tegelik dimensioon võib olla isegi väiksem, olenedes algandmetest. Näiteks saame kaotada ühe dimensiooni, kui andmetabel *N* sisaldab kaht võrdelist rida, kuna sel juhul saame kaks identset reaprofiili (kaks ühtivat punkti pilves). Täpsemalt öeldes saab "õige" dimensioonide arvu pilves kindlaks teha andmemaatriksi *N* astaku kaudu:

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

Samal viisil saame kasutusele võtta ja kirjeldada duaalset pilve ehk veeruprofiilide pilve, mis sisaldab *J* punkti (veeruprofiilid) *I*-dimensionaalses ruumis. Tähistame selle pilve järgnevalt:

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

Pilve $N_A(B)$ elementide *massid* on võrdsed vastavate marginaaltõenäosustega f_j. Taaskord defineeritakse pilve $N_A(B)$ tsentroid f_A, kui tema elementide kaalutud keskmine. Seekord on tsentroid võrdne muutuja *A* marginaaljaotusega:

$$f_A = (f_1, \ldots, f_I).$$

1.2.2 Korrespondentsanalüüsi realiseerimine

Formuleerime korrespondentsanalüüsi eesmärgi. Geomeetrilises mõttes on eesmärk määrata madala dimensiooniga alamruum, mis on 'lähim' kõigile punktidele pilves (Greenacre, 1984). Üldine idee on sama, mis peakomponentide analüüsis, kuid samuti märkame mitmeid erinevusi, kui täpsemalt määratleda termin 'lähim'. Kui punktidel on erinevad massid f_i, siis alamruum asetseb lähemal suurema massiga punktidele, samal ajal kui väiksema massiga punktide kõrvalekalle on kergemini lubatud.

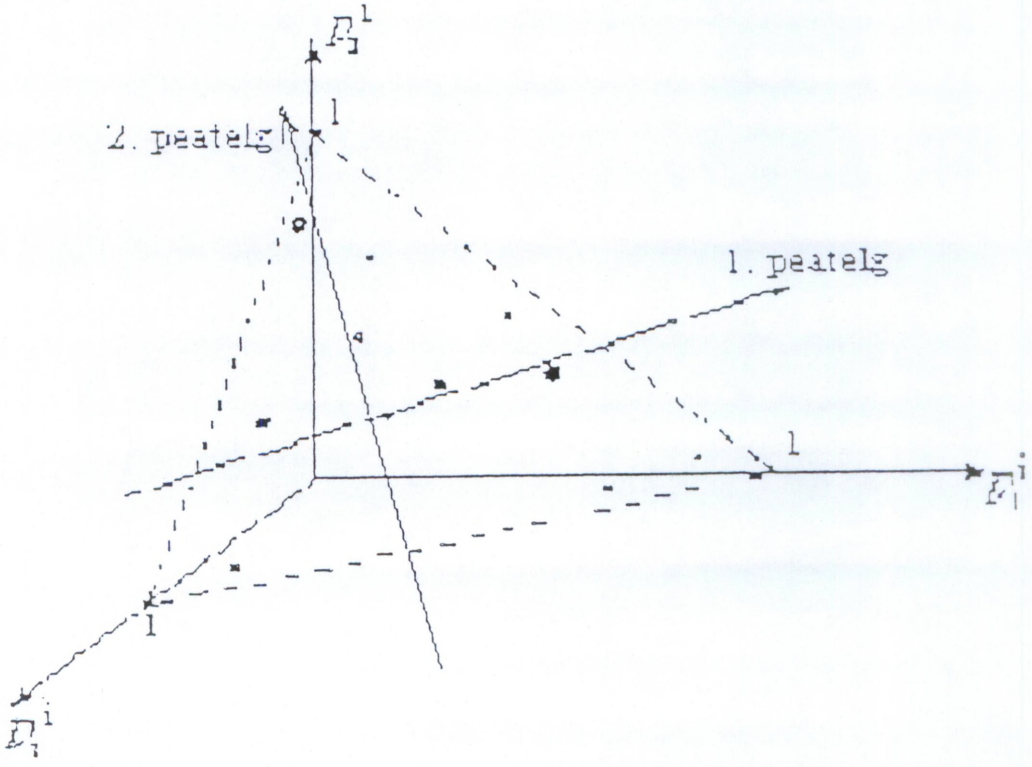
Näide.

Olgu meil andmed koondatud (10×3)-sagedustabelisse. Pärast kümne reaprofiili (tinglikud jaotused veergude järgi) arvutamist näeme, et need kümme punkti asetsevad tegelikult 2-mõõtmelisel simpleksil, sest

$$f_1^i + f_2^i + f_3^i = 1,\ (i = 1,2,\ldots,10) \text{ kõik } f_j^i \geq 0.$$

Kõik need punktid saab kujutada järgneval joonisel.

Joonis 1.1. Kahemõõtmeline simpleks.



Naaseme reaprofiili pilve $N_B(A)$ juurde ja defineerime *kauguse* $d^2(i)$, mis on elemendi f_B^i kaugus pilve tsentroidist

$$d^2(i) = \frac{\sum_j (f_j^i - f_j)^2}{f_j}. \qquad (1.1)$$

Võrdusega (1.1) defineeritud suurust kutsutakse ka 'χ^2-kauguseks', sest see on võrdeline tavalise χ^2-statistikuga, mida kasutatakse testimaks tingliku jaotuse f_j^i ja antud (marginaal) jaotuse f_j erinevust.

Kaugus (1.1) on erinev tavalisest eukleidilisest kaugusest kahe punkti vahel (profiili ja tsentroidi), kuna ta sisaldab normeerivaid faktoreid f_j. Kui me teisendame profiile, siis valem (1.1) on vaadeldav kui tavaline eukleidiline kaugus. Selleks venitame ruumi baasivektori $\sqrt{f_j}$ korda pikemaks, või mis on samaväärne, jagame iga profiili elemendi f_i^j vastava keskmise profiili elemendi ruutjuurega $\sqrt{f_j}$. Seega võime χ^2-kaugust (1.1) vaadata, kui kaalutud eukleidilise kauguse ruutu.

Oluline mõiste korrespondentsanalüüsis on *inerts*, dispersiooni üldistus. Pilve $N_B(A)$ inerts on defineeritud tema punktide kaalutud keskmise kaugusena pilve tsentroidist:

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

kus f_i on i-nda profiili mass.

Inerts on mõõt, mis näitab, kui palju profiilid on hajunud tsentroidi suhtes. Erijuhul, kui kõik massid f_i on võrdsed, muutub inerts I punkti kogudispersiooniks. Inerts on lähedalt seotud χ^2-statistikuga, mis testib sõltumatust kahemõõtmelises sagedustabelis. Nimelt on lihtne näidata, et

$$in(A) = \sum_i \sum_j \frac{(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n})^2}{n_{i\cdot} n_{\cdot j}} =$$

$$= \frac{\chi^2}{n}, \qquad (1.2)$$

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

Antud valem meenutab klassikalist 'Pearsoni χ^2-kordajat', mis on arvutatud sagedustabeli põhjal mõõtmaks seost ridade ja veergude vahel.

Mis puutub duaalpilve $N_A(B)$, siis tema inerts on kindlaks määratud järgneva võrdusega:

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

kus

$$d^2(j) = \sum_i \frac{(f_i^j - f_i)^2}{f_i}$$

on duaalpilve tsentroidi f_B veeruprofiili f^j kaalutud eukleidiline kaugus. Valem (1.2) kehtib samuti duaalse pilve korral, s.t.

$$in(B) = \frac{\chi^2}{n}.$$

Kuna mõlema pilve inertsid on võrdsed, siis nende kahe inertsi ühise väärtuse tähistame λ:

$$\lambda = in(A) = in(B) = \frac{\chi^2}{n}.$$

Korrespondentsanalüüsi eesmärk – leida väiksema dimensiooniga alamruum, mis on lähedal pilve punktidele – määratakse kindlaks inertsi peatelgedega. Peateljed on K-mõõtmelised vektorid, mis lähtuvad pilve tsentroidist ja näitavad suurima inertsi suunda. Esimene peatelg valitakse suunas, mille inerts on maksimaalne. Järgnevad peateljed valitakse nii, et piki vastavat telge inerts oleks maksimaalne ning iga järgnev telg on ortogonaalne kõigi teiste eelnevatega. Saadavas väikese dimensiooniga ruumis kasutatakse praktikas vaid esimesi K^* peatelgi (tavaliselt K^*=2 või 3), lootes, et need K^* dimensiooni kirjeldavad andmeid piisavalt hästi. Praktikas on väga tavaline, et andmeid kujutava pildi saamiseks kasutatakse ainult kaht esimest dimensiooni, kusjuures tulemuseks on kahemõõtmeline joonis rea- ja veeruprofiilidest. Matemaatiline lahendus peatelgede leidmise probleemile saadakse *singulaarväärtuse lahutuse* abil.

1.2.3 Singulaarväärtuse lahutus (SVL)

Olgu X $(m\times n)$-maatriks. Siis maatriksi X^TX omaväärtuste λ_i ruutjuuri $\alpha_i=\sqrt{\lambda_i}$ nimetatakse maatriksi X *singulaarväärtusteks*: $\alpha_1 \geq \ldots \geq \alpha_n$.

Eeldame, et maatriks $X_{[m\times n]}$ on astakuga K. Lahutust

$$X = U\Delta_\alpha V^T \tag{1.3}$$

nimetatakse *singulaarväärtuse lahutuseks*, kui Δ_α on diagonaalmaatriks ning U ja V on maatriksid, mille veeruvektorid on ortogonaalsed, kusjuures $[K\times K]$-diagonaalmaatriksi Δ_α diagonaalelemendid on maatriksi X esimesed K mittenegatiivset singulaarväärtust: $\alpha_1 \geq \ldots \geq \alpha_K > 0$, U on $(m\times K)$ ja V on $(n\times K)$ ja $U^TU = V^TV = I_K$.

Maatriksi U veerge nimetatakse maatriksi X *vasakpoolseteks singulaarvektoriteks*. Need on ortonormeeritud baasiks maatriksi X veergude jaoks ning ühtlasi on nad maatriksi XX^T omavektoriteks, mis vastavad XX^T omaväärtustele.

Maatriksi V veerge nimetatakse maatriksi X *parempoolseteks singulaarvektoriteks*. Need on ortonormeeritud baasiks maatriksi X (transponeeritud) ridade jaoks ning ühtlasi on nad maatriksi X^TX omavektoriteks, mille nullist erinevad omaväärtused ühtivad maatriksi XX^T nullist erinevate omaväärtustega. (Lay, 1994).

1.2.4 Üldistatud singulaarne lahutus

Korrespondentsanalüüsis vajame singulaarväärtuse lahutuse üldistust.
Olgu $\Omega_{[I\times I]}$ ja $\Phi_{[J\times J]}$ on positiivselt määratud sümmeetrilised maatriksid. Esitame tavalise singulaarse lahutuse kujul (1.3) maatriksile $\Omega^{1/2}X\Phi^{1/2}$:

$$\Omega^{1/2}X\Phi^{1/2} = U\Delta_\alpha V^T,$$

kus $U^TU = V^TV = \mathbf{I}_K$ ja Δ_α on üldistatud singulaarväärtuste diagonaalmaatriks, kus singulaarväärtused on järjestatud mittekahanevalt.

Siinjuures maatriksite $A^{\frac{1}{2}}$ on määratud sümmeetrilise maatriksi korral võrdusega $A = A^{\frac{1}{2}}A^{\frac{1}{2}}$. Valides

$$L \equiv \Omega^{-1/2}U\,,\ M \equiv \Phi^{-1/2}V\,,$$

saame üldistatud SVL maatriksi X jaoks:

$$X = L\Delta_\alpha M^T\,, \tag{1.4}$$

kus $L^T\Omega L = I_K$ ja $M^T\Phi M = I_K$.

Maatriksite L ja M veerud on ortonormeeritud üldistatud (kaalutud) eukleidiliste meetrikate Ω ja Φ suhtes, vastavalt. Need on üldistatud vasakpoolseteks ja parempoolseteks singulaarvektoriteks ning ühtlasi on nad ortonormeeritud baasiks maatriksi X veergude ja ridade jaoks.

Alternatiivselt võime kirjutada valemi (1.4) kujul

$$X = \sum_{k=1}^{K} \alpha_k l_k m_k^T\,,$$

kus l_k ja m_k tähistavad maatriksite L ja M k-ndaid veerge vastavalt.

1.2.4. Singulaarväärtuse lahutus korrespondentsanalüüsis

Anname lühiülevaate üldistatud singulaarväärtuse lahutuse kasutamisest korrespondentsanalüüsis.

Toome sisse täiendavad tähistused. Olgu R ja C andmemaatriksi N rea- ja veerusummade diagonaalmaatriksid:

$$R = diag(n_{1\cdot},\ldots,n_{I\cdot})$$

ja

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

Olgu $m_1,\ldots,m_K$ inertsi peateljed reaprofiili pilve jaoks. Praktika jaoks on oluline teada pilve elementide koordinaate nende baasivektorite suhtes. Olgu x_{ik} reaprofiili koordinaat k-nda inertsi peatelje suhtes (nimetame *peakoordinaadiks*). Siis i-s reaprofiil on iseloomustatud K-vektori $x_1 = (x_{i1}, x_{i2}, \ldots, x_{iK})$ poolt ja kogu reaprofiili pilv on iseloomustatud $(I \times K)$-maatriksi X poolt,

$$X = \begin{bmatrix} \underline{x_1} \\ \underline{x_2} \\ \underline{\cdots} \\ \underline{x_I} \end{bmatrix}.$$

Singulaarväärtuse lahutuse meetodist tuleneb, et maatriksi X veerud peavad olema $(I \times I)$-maatriksi $R^{-1}N\ C^{-1}N^T$ omavektorid, mis vastavad $R^{-1}NC^{-1}N^T$ mittenullilistele omaväärtustele $\lambda_1,\ldots,\lambda_K$. See tähendab, et maatriks X rahuldab võrrandit

$$(R^{-1}N\ C^{-1}N^T)X = XD_\lambda, \qquad (1.5)$$

kus D_λ on omaväärtustest koosnev $(K \times K)$-diagonaalmaatriks, $D_\lambda = (\lambda_1,\ldots,\lambda_K)$ ja eeldame, et $\lambda_2 \geq,\ldots,\geq \lambda_K > 0$. Tegelikult maatriksi $R^{-1}N\ C^{-1}N^T$ mittenegatiivsete omaväärtuste arv on $rank(N) = K+1$. Välja on jäetud 'triviaalne' suurim omaväärtus $\lambda_0 = 1$ ja vastav konstantne omavektor, mis koosneb ühtedest ja mis rahuldab triviaalselt valemit (1.5). Geomeetriliselt tähendab triviaalse omaväärtuse ja omavektori ärajätmine analüüsis seda, et me asetame peatelgede alguspunkti pilve tsentroidile.

Selleks, et saada ühest lahendit omaväärtusprobleemile (1.5), on vaja fikseerida omavektori (veeru X) pikkus. Mõistlik on kasutada standardiseerimist:

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

mis määrab punktide koordinaatide kaalutud ruutude summa (s.t. kaalutud dispersiooni ehk inertsi) piki k-ndat peatelge pilves on võrdne omaväärtusega λ_k ja me nimetame seda väärtust k-ndaks *peainertsiks*.

Fakti, et reaprofiili tsentroid asub peatelje alguspunktis, saab väljendada järgnevalt

$$\sum_i n_{i\cdot} x_i = 0. \qquad (1.6)$$

Lihtne järeldus on siit, et mõni profiili koordinaat peab olema negatiivne suvalise dimensiooni korral.

Korrespondentsanalüüsi rakendustes on küllaltki tavaline, et arvesse võetakse ainult kaks esimest inertsitelge. See tähendab, et tavaliselt kasutatakse vaid kaht esimest veergu maatriksist X saamaks esimest ja teist koordinaati iga reaprofiili jaoks. Neid koordinaate kasutatakse tavalisel viisil, et kujutada reaprofiile I punktina tasandil.

Duaalpilv koosneb veeruprofiilidest, sel juhul tähistame inertsi peateljed $l_1,\ldots,l_K$ ja vastavad j-ndad veeruprofiili koordinaadid $y_{j1}, y_{j2},\ldots, y_{jK}$. Viimased saab koondada $(J \times K)$-maatriksisse Y. Singulaarväärtuse lahutuse kohaselt peab Y rahuldama võrrandit

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

kus maatriks D_λ on sama, mis valemis (1.3). Sobiv standardiseering on nüüd

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

mis annab sama peainertsi piki iga telge nagu ka reaprofiilide pilve korral. Valemi (1.6) asemel kehtib meil nüüd võrdus maatriksi Y ridade y_j jaoks:

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

Võrdsus mõlema pilve peainertsis võimaldab ühildada kaks erinevat graafilist pilti ühiseks pildiks, kus on I reaprofiili punkti ja J veeruprofiili punkti. Vastavalt Greenacre (1984) soovitusele peaksime vältima ohtu interpreteerida kaugusi erinevate pilvede vahel, kuna sellist kaugust ei ole otseselt defineeritud. Põhjus on selles, et me kasutame kahte erinevat baasi kahe pilve jaoks, seetõttu ei ole koordinaadid otseselt võrreldavad. Selle asemel on olemas lihtne seos peakoordinaatide vahel:

$$YD_\rho = C^{-1}N^T X, \qquad (1.7)$$

$$XD_\rho = R^{-1}NY, \qquad (1.8)$$

kus $D_\rho = D_\lambda^{\frac{1}{2}}$ - diagonaalmaatriks mille peadiagonaalil on $\rho_k = \sqrt{\lambda_k}$.

Kasulik on kirjutada valemid (1.7) ja (1.8) elementide kaupa:

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

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

Esimene võrdus näitab, et iga veerukoordinaat on reakoordinaatide kaalutud keskmine (kuni konstantse teguri ρ_k täpsuseni), mille kaalud võrduvad lahtrite arvuga selles veerus. Teine võrrand näitab, et iga reakoordinaat on veerukoordinaatide kaalutud keskmine (kuni konstantse teguri ρ_k täpsuseni), kasutades kaale, mis võrduvad lahtrite arvuga selles reas. Sellist operatsiooni kutsutakse *pöördkeskmistamiseks* ja seda kasutatakse tihti arvutamaks peakoordinaate, eriti kui on tegemist suurte andmetabelitega. Vahel kasutatakse seda terminit ka märkimaks korrespondentsanalüüsi ennast.

Lõpuks selgitame omaväärtuste $\lambda_1, \lambda_2, \ldots, \lambda_K$ tähendust, näidates, et nad määravad tegelikult kogu inertsi λ osad. Meenutame esiteks, et pilve koguinerts $in(A) = in(B) = \frac{\chi^2}{n}$ on pilve tsentroidi ümber paiknevate pilve punktide hajuvuse mõõt. Selle mõõdu saab jagada *K* osaks järgnevalt:

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

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

$$= n(\lambda_1 + \lambda_2 + \ldots + \lambda_K), \qquad (1.9)$$

kuna maatriksi jälg $tr(R^{-1} N C^{-1} N^T) = \sum_{i=1}^{I} \lambda_i$, aga kuna *K=I-1* ja triviaalne suurim omaväärtus $\lambda_0 = 1$, siis $\chi^2 = n(\lambda_1 + \cdots + \lambda_K)$.

Näitame, et $tr(R^{-1}NC^{-1}N^T) = \sum_i \sum_j \frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}$. Avaldame kõigepealt peadiagonaali üldelemendi:

$$(R^{-1}NC^{-1}N^T)_{ii} = \sum_{j=1}^{J} \frac{1}{n_{i\cdot}} n_{ij} \frac{1}{n_{\cdot j}} n_{ij} = \sum_{j=1}^{J} \frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}.$$

Seega

$$tr(R^{-1}NC^{-1}N^T) = \sum_i \sum_j \frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}},$$

mida oligi vaja näidata.

Inertsi avaldisest järeldub, et

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

Seega on koguinerts jaotatud K peatelje vahel, esimene telg kirjeldab suurima osa inertsist jne. See on kogu inertsi lahutusvalem analoogne dispersiooni lahutusele. Samal ajal valem (1.9) annab χ^2-statistiku lahutuse sagedustabelis. Arvuti väljatrükid annavad tavaliselt suhtelise inertsi $\frac{\lambda_k}{\lambda}$ (väljendatuna protsentides) näitamaks erinevate telgede suhtelist tähtsust andmete kirjeldamisel.

2. Korrespondentsanalüüsi rakendamine

2.1 Andmed

Korrespondentsanalüüsi rakendamisel kasutame semestritöö (Koov, 2004) andmeid. Semestritöös tegime statistilise ülevaate Tartu Ülikooli õppejõududest ja teaduritest.

Töös kasutatavad Tartu Ülikooli õppejõudude ja teadustöötajate, edaspidi TÜ töötajate, andmed on saadud Tartu Ülikooli personaliosakonnast (aprill, 2003). Saadud andmestik sisaldas seitset tunnust ja 1212 inimest. Aga kuna andmestik sisaldas ka lünki ja kordusi, siis kõigepealt püüdsime neid kõrvaldada. Peamised kordused seisnesid selles, et mõni TÜ töötaja oli andmestikus esindatud kaks korda, samas oli ka puuduvaid tunnuste väärtusi. Kõige sagedamini puudusid tunnuste "sünnikoht" ja "teaduskraad" väärtused. Lõpptulemusena eemaldasime andmestikust 111 inimest, kelle kohta ei õnnestunud täiendavaid andmeid saada. Peamiselt, olid nad kas assistendid, õpetajad või teadurid. Suurem osa kõrvaldatuid olid naissoost ning peamiselt olid nad filosoofiateaduskonna või haridusteaduskonna töötajad. Analüüsiks kasutatav andmestik sisaldab 1101 kirjet ning seitset tunnust. Need vaadeldavad seitse tunnust on:

- Nimi
- Sugu
- Sünnikoht
- Vanus
- Teaduskraad
- Ametinimetus
- Teaduskond

2.2 Kodeerimine

Andmestik, mis saadi Tartu Ülikooli personaliosakonnast, ei sobinud sellisel kujul statistilise analüüsi tegemiseks. Selleks, et läbi viia statistilist andmete analüüsi ja ka korrespondentsanalüüsi, oli esmalt vaja tunnused korrastada ja kodeerida.

Tunnused "nimi" ja "vanus".

Antud andmestiku puhul ei vajanud kodeerimist tunnused "nimi" ja "vanus". Tunnust "nimi" otseselt analüüsis ei kasutata, ta on vaid identifitseerivaks tunnuseks. Ja tunnus "vanus" on antud juba eluaastates, mis ei vaja ümberkodeerimist. Tunnuse "vanus" jagasime rühmadesse kümne aasta kaupa:

1. ≤ 29
2. 30-39
3. 40-49
4. 50-59
5. ≥ 60

Tunnus "sugu".

Tunnus "sugu" oli antud andmestikus kas M (mees) või N (naine). Selle tunnuse kodeerisime vastavalt 1 – mees ja 2 – naine.

Tunnus "sünnikoht".

Tunnus "sünnikoht" vajas enne kodeerimist korrastamist, sest saadud andmestikus olid sünnikohad märgitud väga paljude asulate ja linnade nimetustena. Kodeerisime tunnuse maakondade järgi ning eraldi kaks suuremat linna - Tartu ja Tallinn:

1. Tartu linn
2. Tallinn
3. Hiiumaa
4. Läänemaa
5. Harjumaa
6. Lääne-Virumaa
7. Ida-Virumaa
8. Raplamaa

9. Järvamaa
10. Jõgevamaa
11. Saaremaa
12. Pärnumaa
13. Viljandimaa
14. Tartumaa
15. Valgamaa
16. Põlvamaa
17. Võrumaa

Välisriigid kodeerisime järgnevalt:

18. Vene
19. Välismaa

"Vene" alla paigutasime endise NSVL territooriumi väljaspool Eestit, "välismaa" alla aga kõik ülejäänud riigid.

Tunnus "teaduskraad".

Tunnuse "teaduskraad" kodeerisime kolmeks erinevaks rühmaks, mis oleksid järgnevad:

1. Doktor – Dok
2. Magister – Mag
3. Kraadita - Krd

Kood "Dok" on omistatud kõigile isikutele, kellel on doktorikraad või kes on teaduste kandidaadid. Kood "Mag" omistasime isikutele, kes omavad magistrikraadi või on lõpetanud arstiteaduskonna. Kõik ülejäänud TÜ töötajad kodeeriti kraadita töötajatena.

Tunnus "teaduskond".

Kokku on Tartu Ülikoolis üksteist teaduskonda ning lisaks võtsin kokku ühe koodi alla Tartu Ülikooli iseseisvad instituudid nagu TÜ Eesti mereinstituut, TÜ tehnoloogiainstituut ja TÜ õigusinstituut.

Tunnuse "teaduskond" kodeerimine oli järgmine:

1. USUS – Usuteaduskond
2. OIOI – Õigusteaduskond
3. ARAR – Arstiteaduskond
4. FLFL – Filosoofiateaduskond
5. BGBG – Bioloogia-geograafiateaduskond
6. FKFK – Füüsika-keemiateaduskond
7. HTHT – Haridusteaduskond
8. KKKK – Kehakultuuriteaduskond
9. MJMJ – Majandusteaduskond
10. MTMT – Matemaatika-informaatikateaduskond
11. SOSO – Sotsiaalteaduskond
12. Instituut – TÜ instituudid

Tunnus "ametinimetus".

Tunnuses "ametinimetus" on kokku võetud Tartu Ülikoolis töötavate inimeste ametinimetused. Need jagasime nelja klassi: professor, dotsent, teadur, assistent. "Assistendi" alla kuuluvad kraadita assistendid ja õpetajad. Kodeeritud tunnuse "teadur" alla kuuluvad need isikud, kelle ametinimetuseks on kas teadur, lektor või magistrikraadiga assistent. Koodi "dotsent" said lisaks dotsentidele ka vanemteadurid. Seega loetelu sai järgmine:

1. Professor
2. Dotsent
3. Teadur
4. Assistent

2.3 Korrespondentsanalüüsi tulemused

Analüüsimaks andmeid korrespondentsanalüüsi meetodiga kasutame paketti SAS. Selleks kasutame SAS'is olevat korrespondentsanalüüsi protseduuri CORRESP. Siinkohal toome välja protseduuri süntaksi, mis on järgnev:

PROC CORRESP <options>;

TABLES <row-variables,> <column-variables>;

VAR variables;

ID variables;

SUPPLEMENTARY variables;

WEIGHT variables;

RUN;

%plotit (data=faili nimi, datatype=corresp)

Esimesel real tuleb ära määrata andmestik, mille põhjal korrespondentsanalüüsi tehakse, käsuga DATA= andmestiku nimi ning ka väljundandmestik, mille põhjal tehakse interpreteeritav graafik (süntaksi viimase lausega), selleks on käsk OUTC= väljundfaili nimi.

Sõltuvalt sellest, kas andmed on algfailis esitatud sagedustabelina või objekt-tunnus maatriksina, tuleb valida andmete kujutamiseks käskude komplekt, kas TABLES või VAR.

Kui andmed on esitatud sagedustabelina, siis kasutatakse võtmesõna VAR, mille järel loetletakse komadega eraldatult tunnuste nimed, mida soovitakse analüüsis kasutada kui sagedustabeli veerge. Loodava sagedustabeli reatunnus sisestatakse võtmesõna ID järel. Käsuga VAR märgitud tunnuste väärtused peavad olema mittenegatiivsed ja arvulised, kuna tegemist on sagedustabeliga.

Kui andmestik on esitatud objekt-tunnus maatriksina, siis kasutatakse käsku TABLES koos tunnuste nimedega, mis määravad sagedustabeli. Tunnuste sisestamisel peab kõigepealt sisestama reatunnuse või tunnused ja seejärel veerutunnused; reatunnused eraldatakse veerutunnustest komaga.

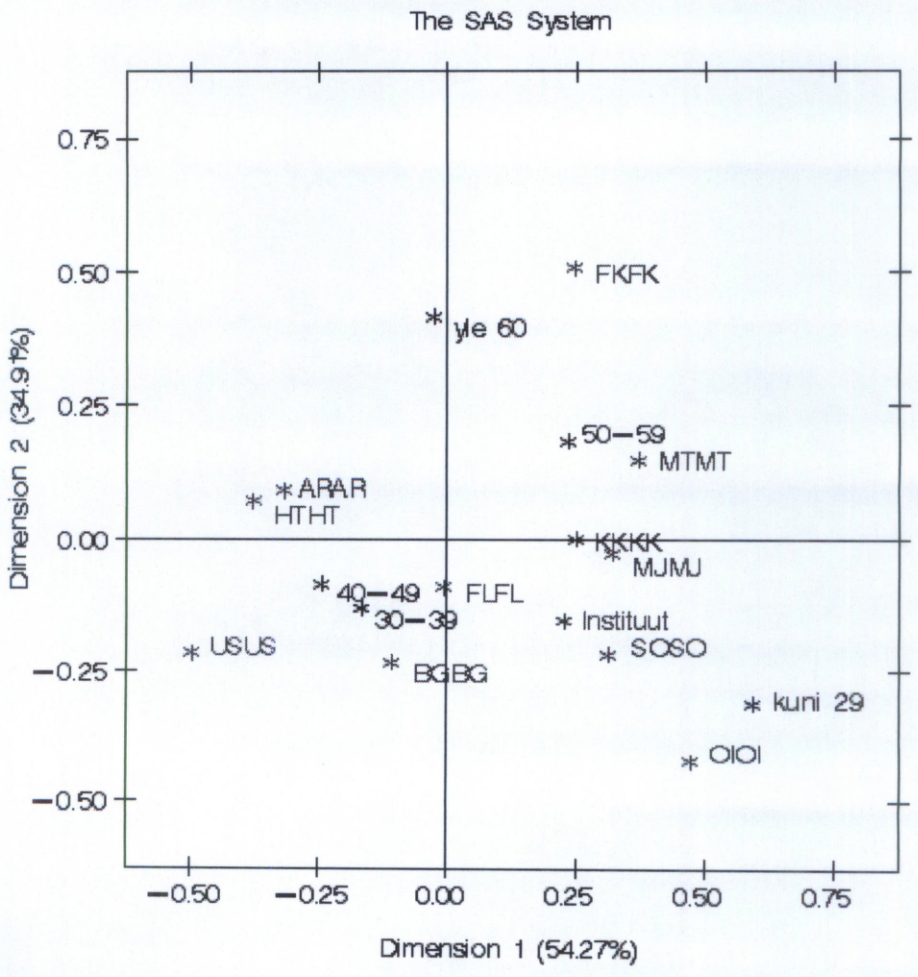
Koos TABLES või VAR lausega võib kirja panna ka lause SUPPLEMENTARY, millega saab lisada tunnused, mis pannakse tehtavale joonisele, kuid mida ei kasutata arvutuste tegemisel. Võimalik on kasutada ka kaalutunnust WEIGHT erinevatele väärtustele erinevate kaalude andmiseks.

Rakendame nüüd SAS'i korrespondentsanalüüsi TÜ õppejõudude ja teadurite andmestikule. Kõigepealt uurime, millised on seosed tunnuste 'teaduskond' ja 'vanusgrupp' vahel. Selle põhjal saab vaadelda, millised teaduskonnad on "vanad" ja millised "noored" oma töötajaskonna poolest. Selleks, et tulemusi interpreteerida, esitame joonise ja olulisemad tulemused. (Lisa 1, Tabel 1, lk. 35-36).

Tabel 2.1. Korrespondentsanalüüsi tulemused.

Dimensioon	Inerts	Kirjeldavuse protsent	Kumulatiivne kirjeldatus
1	0.07280	54.27	54.27
2	0.04684	34.91	89.19
3	0.01189	8.86	98.05
4	0.00262	1.95	100.00
Kokku	0.13414	100.00	

Joonis 2.1. Tunnuste 'teaduskond' ja 'vanusgrupp' kahemõõtmeline joonis.



Tabelist 2.1 on näha, et koguinerts on 0.13414, millest esimene telg kirjeldab 0.0728 ehk 54.27% ja teine telg kirjeldab 0.04684 ehk 34.91%. Esimene ja teine telg kokku kirjeldavad 89.19% koguinertsist. Vertikaaltelge võib interpreteerida, kui vanusetelge, kus all on nooremad ja üleval vanemad töötajad. Jooniselt näeme ka, et vanusegrupid 40-49 ja 30-39 asuvad lähestikku. "Noorim" teaduskond on õigusteaduskond ka võib öelda, et sotsiaalteaduskond on "noor", aga seal on ka palju vanemaid TÜ töötajaid, seepärast ta asubki kaugemal 'kuni 29'-aastastest. Suurem osa teaduskondi kuulub keskmistesse vanusegruppidesse.

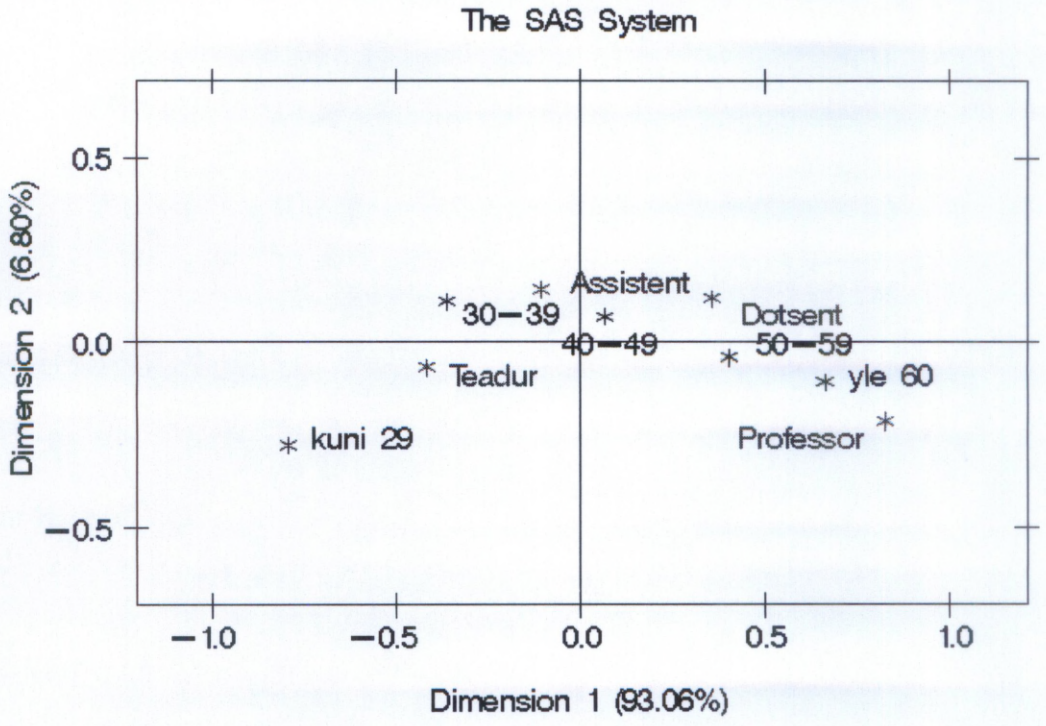
Joonise põhjal võib öelda, et "vanim" teaduskond on füüsika-keemiateaduskond, selle teaduskonna töötajad on suuremalt osalt üle 50 aastased. Näeme ka, et matemaatika-informaatikateaduskond kuulub "vanemate" teaduskondade hulka.

Vaatame nüüd tunnuste 'ametinimetus' ja 'vanusgrupp' korrespondentsanalüüsi tulemusi. Tulemused on toodud lisas (Lisa 1, Tabel 2. lk. 37-39) ning olulisemad arvud alljärgnevas tabelis.

Tabel 2.2. Korrespondentsanalüüsi tulemused.

Dimensioon	Inerts	Kirjeldavuse protsent	Kumulatiivne kirjeldatus
1	0.20002	93.06	93.06
2	0.01461	6.80	99.86
3	0.00030	0.14	100.00
Kokku	0.21493	100.00	

Joonis 2.2. Tunnuste 'ametinimetus' ja 'vanusgrupp' kahemõõtmeline joonis.



Saadud joonise kohta saame öelda, et punktide paiknemine on paraboolse kujuga. Horisontaalne peatelg on antud joonise korral interpreteeritav kui vanuseline telg. Tabelist 2.2 näeme, et kaks esimest telge kirjeldavad kokku 99.86% koguinertsist. Suurema osa kirjeldab esimene peatelg nimelt 93.06% ja teine telg kirjeldab vaid 6.8%.

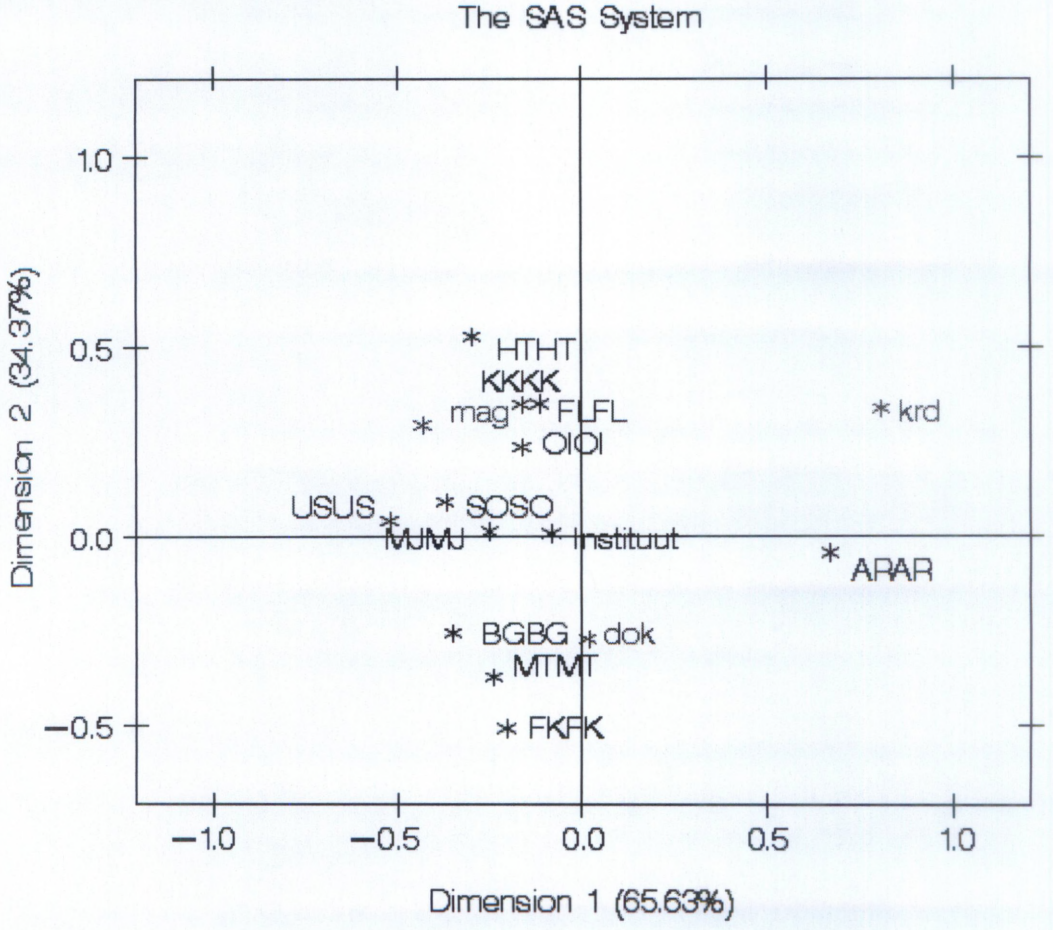
Jooniselt näeme, et iga ametinimetus kuulub enamasti kahte vanusegruppi. Teadurid kuuluvad enamasti vanusegruppi 'kuni 39', assistendid vanusegruppi '30-49', dotsendid vanusegruppi '40-59' ja professorid vanusegruppi 'üle 50'. Selline jaotus on suhteliselt loogiline. Selleks et TÜ töötaja omaks dotsendi või professori ametinimetus peab ta olema selleks juba palju aastaid töötanud ja seepärast ongi vanemad isikud sellistel ametikohtadel.

Uurime ka tunnuste 'teaduskraad' ja 'teaduskond' vahelisi seoseid (Lisa 1, Tabel 3, lk. 39-40). Eesmärgiks on välja selgitada, mis teaduskraad on eri teaduskondades domineeriv. Selleks uurime järgnevat tabelit ja joonist.

Tabel 2.3. Korrespondentsanalüüsi tulemused.

Dimensioon	Inerts	Kirjeldavuse protsent	Kumulatiivne kirjeldatus
1	0.15858	65.63	65.63
2	0.08304	34.37	100.00
Kokku	0.24163	100.00	

Joonis 2.3. Tunnuste 'teaduskraad' ja 'teaduskond' kahemõõtmeline joonis.



Esimene telg kirjeldab 0.15858 ehk 65.63% ja teine telg 0.08304 ehk 34.37% koguinertsist. Ja kokku kirjeldavad mõlemad teljed 100%, see tähendab, et kõik andmed saavad kirjeldatud.

Jooniselt näeme, et kraadita isikud jäävad kõigist teaduskondadest väga kaugele. Kõige lähema on arstiteaduskond, kus on ka kõige enam kraadita töötajaid, aga samas ei ole arstiteaduskond väga kaugel doktorikraadist, võiks öelda, et nad on peaaegu ühekaugusel. Kõige enam on doktorikraadiga töötajaid bioloogia-geograafiateaduskonnas, matemaatika-informaatikateaduskonnas ja füüsika-

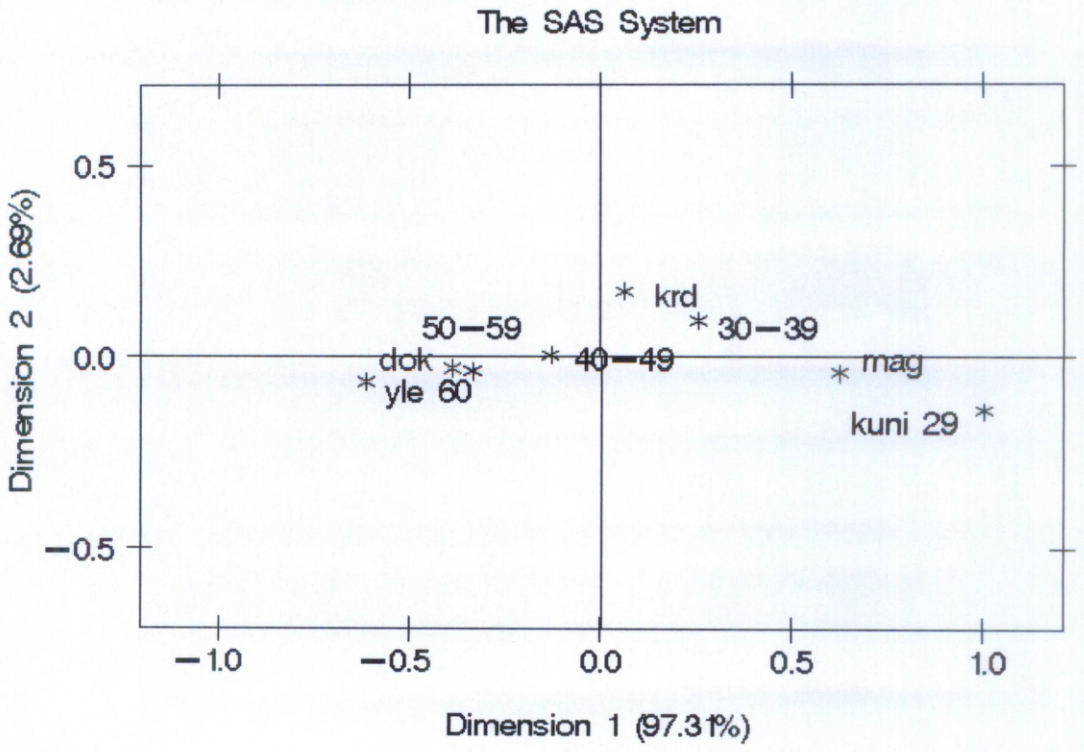
keemiateaduskonnas. Ülejäänud teaduskondades domineerib magistrikraad. Samas võime öelda, et majandusteaduskonnas, usuteaduskonnas ja sotsiaalteaduskonnas on ilmselt doktorikraadi ja magistrikraadi omanikke suhteliselt võrdselt. Graafiliselt asetsevad need teaduskonnad doktorikraadi ja magistrikraadi vahel.

Viime korrespondentsanalüüsi läbi ka tunnustega 'teaduskraad' ja 'vanusgrupp' (Lisa 1, Tabel 4, lk. 41-42). On ootuspärane, et vanemad isikud omavad doktorikraadi ja nooremad magistrikraadi.

Tabel 2.4. Korrespondentsanalüüsi tulemused.

Dimensioon	Inerts	Kirjeldavuse protsent	Kumulatiivne kirjeldatus
1	0.20307	97.31	97.31
2	0.00560	2.69	100.00
Kokku	0.20867	100.00	

Joonis 2.4. Tunnuste 'teaduskraad' ja 'vanusgrupp' kahemõõtmeline joonis.



Esimene telg kirjeldab suurema osa koguinertsist, nimelt 97.31% ja teine telg vaid 2.69%. Horisontaaltelge võib nimetada vanuseliseks teljeks, mille paremal pool on

noorim vanusegrupp ja vasakul pool vanim vanusegrupp. Kui vaadata joonisel doktorikraadi ja magistrikraadi, siis on näha, et tekib kaks rühma. Magistrikraadi juurde kuuluvad vanusegrupid '30-39' ja 'kuni 29' ning doktorikraadi juurde kuuluvad vanusegrupid '40-49', '50-59' ja 'üle 60'. Seega ongi nii, et nooremad isikud omavad magistrikraadi ja vanemad isikud doktorikraadi. Isikud, kellel pole kraadi, on enamasti vanuses '30-39'.

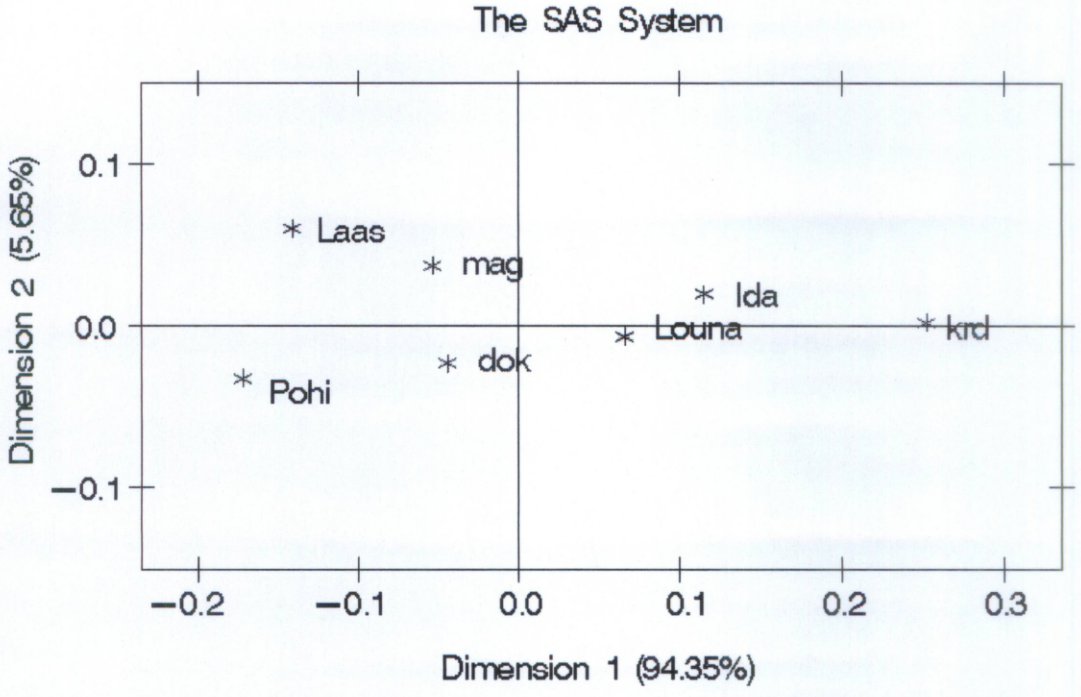
Uurime ka veel tunnuse 'sünnikoht' seoseid teiste tunnustega. Parema ülevaate saamiseks jaotame Eesti maakonnad Põhja-, Lõuna-, Lääne- ja Ida-Eestiks järgnevalt:
Põhja-Eesti – Tallinn, Harjumaa;
Lõuna-Eesti – Tartu, Tartumaa, Viljandimaa, Jõgevamaa, Põlvamaa, Valgamaa, Võrumaa;
Lääne-Eesti – Läänemaa, Hiiumaa, Saaremaa, Raplamaa, Pärnumaa;
Ida-Eesti – Lääne-Virumaa, Ida-Virumaa, Järvamaa.

Uurime tunnuste 'sünnikoht' ja 'teaduskraad' vahelisi seoseid (Lisa 1, Tabel 5, lk. 43-44). Selleks uurime järgnevat tabelit ja joonist.

Tabel 2.5. Korrespondentsanalüüsi tulemused.

Dimensioon	Inerts	Kirjeldavuse protsent	Kumulatiivne kirjeldatus
1	0.01213	94.35	94.35
2	0.00073	5.65	100.00
Kokku	0.01286	100.00	

Joonis 2.5. Tunnuste 'teaduskraad' ja 'sünnikoht' kahemõõtmeline joonis.



Tabelist 2.5 näeme, et koguinerts on 0.01286, millest esimene telg kirjeldab 0.01213 ehk 94.35% ja teine telg 0.00073 ehk 5.65%.

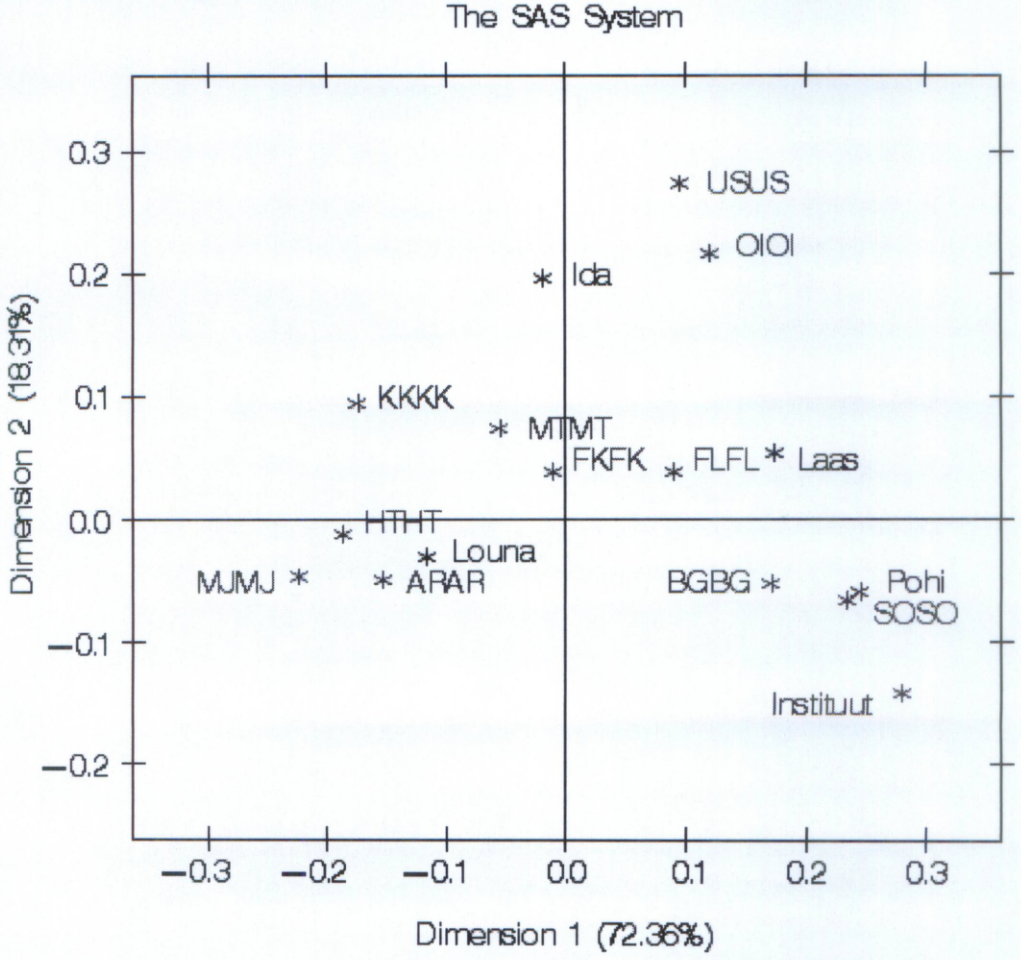
Joonist interpreteerides võime öelda, et kraadita töötajad on enamasti pärit, kas 'Lõunast' või 'Idast'. Vaadates, aga doktorikraadi ja magistrikraadi, siis need teaduskraadid asuvad sünnipiirkondadest suhteliselt võrdsetel kaugustel. Seega võime väita, et teaduskraade kaitstakse olenemata sünnikohast. Teaduskraade kaitsevad siiski TÜ töötajad, kelle on huvi teadustöö vastu.

Rakendame veel korrespondentsanalüüsi tunnustele 'teaduskond' ja 'sünnikoht' (Lisa 1, Tabel 6, lk. 45-46). Analüüsime, kas mõned teaduskonnad on tugevamini seotud teatud piirkondadega.

Tabel 2.6. Korrespondentsanalüüsi tulemused.

Dimensioon	Inerts	Kirjeldavuse protsent	Kumulatiivne kirjeldatus
1	0.02302	72.36	72.36
2	0.00583	18.31	90.67
3	0.00297	9.33	100.00
Kokku	0.03182	100.00	

Joonis 2.5. Tunnuste 'teaduskond' ja 'sünnikoht' kahemõõtmeline joonis.



Tabelist 2.6 näeme, et teljed kirjeldavad kokku 90.67% koguinertsist. Esimene telg kirjeldab suurema osa, nimelt 72.36% ja teine telg kirjeldab 18.31%.

Saadud pilt ei ole küll kergesti interpreteeritav, aga üht-teist saame siit välja lugeda. Üldiselt on 'Lõuna' kõige suurema massiga, seepärast asuvadki paljud teaduskonnad 'Lõuna' läheduses. Bioloogia-geograafiateaduskond on lähemal 'Põhjale', arvatavasti seetõttu, et mujal ei ole võimalusi seda eriala õppida ja praktiseerida. Arstiteaduskond

asetseb 'Lõuna' lähedal tõenäoliselt perekondliku traditsiooni tõttu s.t., et kui perest keegi on juba arst, siis seda traditsiooni jätkatakse ka järgnevates põlvkondades. Instituudid asetsevad 'Põhja' lähedal, sest TÜ kaks instituuti asuvad Tallinnas. Fakti, et matemaatika-informaatikateaduskond ja füüsika-keemiateaduskond asuvad suhteliselt ühekaugusel eri piirkondadest võib tõlgendada niiviisi, et täppisteaduste esindajad on tulnud ülikooli olenemata sünnikohast.

Kokkuvõte

Uurijatel on tihti vaja analüüsida suuri andmetabeleid, aga neid on raske kujutada hajuvusdiagrammidel ning seetõttu ka raske interpreteerida. Sellisel juhul on abiks korrespondentsanalüüs. Korrespondentsanalüüsiga saab suuri sagedustabeleid või objekt-tunnus maatrikseid kujutada kahemõõtmelisel või kolmemõõtmelisel graafikul, mis on kergesti interpreteeritavad.

Käesolevas töös on antud ülevaade korrespondentsanalüüsi põhimõistetest, tähistustest ja meetodi rakendamisest. On tutvustatud singulaarväärtuse lahutust, mida kasutatakse korrespondentsanalüüsi matemaatiliseks esitamiseks.

Töö teises osas on rakendatud korrespondentsanalüüsi Tartu Ülikooli töötajate andmestikule ning saadud arusaadavad ja interpreteeritavad tulemused.

Korrespondentsanalüüs on kasulik andmeanalüüsi meetod ning seda kasutatakse tänapäeval väga mitmetes valdkondades. Eriti oluline on see meetod mittearvulisi tunnuseid sisaldavate andmestike korral.

Correspondence analysis of the staff of the University of Tartu

Summary

Correspondence analysis is an exploratory method for analyzing large data tables. Rows and columns of the data matrix are presented as points in low-dimensional vector spaces. Usually correspondence analysis gives us two-dimensional or three-dimensional graphs which are easy to interpret. Usually if we have large data tables then it is hard to present data on a scatter plot and so the data is not easily interpretable.

In the paper we introduce basic concept and notation of correspondence analysis and essentials of the mathematical theory that stands behind the correspondence analysis method.

First we give an overview of the correspondence analysis then we explain the basic concept and introduce notation. Then we explain how the correspondence analysis works. Also we give an overview of the mathematical theory of the singular value decomposition which is used in correspondence analysis.

Finally we give an application of correspondence analysis. We use data about teaching staff and researchers of the University of Tartu. Different variables were analyzed by correspondence analysis and the graphical results interpreted.

Nowadays correspondence analysis is used in different areas for studying categorical data. Correspondence analysis is very important method for analyzing large data tables.

Kasutatud kirjandus

1. Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis.* Academic Press, London.
2. Koov, K (2004), *Õppejõud ja teadurid Tartu Ülikoolis: statistiline ülevaade.* Semestritöö, TÜ matemaatilise statistika instituut, Tartu.
3. Lay, D.C. (1994). *Linear Algebra and its Applications.* Addison-Wesley, Reading.
4. Pärna, K. *Correspondence Analysis: an Introduction and Some Examples. Research Report 1993:7*, Stockholm University.

Lisa 1 SAS'i väljatrükid

Tabel 1. Korrespondentsanalüüsi tulemused tunnustele 'teaduskond' ja 'vanusgrupp'.

The CORRESP Procedure
Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	11 22 33 44 55
0.26982	0.07280	80.155	54.27	54.27	**********
0.21642	0.04684	51.567	34.91	89.19	**********
0.10902	0.01189	13.087	8.86	98.05	****
0.05119	0.00262	2.885	1.95	100.00	*
Total	0.13414	147.693	100.00		

Degrees of Freedom = 44

Row Coordinates

	Dim1	Dim2
ARAR	-0.3186	0.0923
BGBG	-0.1073	-0.2327
FKFK	0.2484	0.5081
FLFL	-0.0070	-0.0909
HTHT	-0.3783	0.0721
Instituut	0.2270	-0.1539
KKKK	0.2504	0.0005
MJMJ	0.3214	-0.0275
MTMT	0.3732	0.1473
OIOI	0.4733	-0.4204
SOSO	0.3142	-0.2174
USUS	-0.4981	-0.2145

Summary Statistics for the Row Points

	Quality	Mass	Inertia
ARAR	0.9418	0.2516	0.2191
BGBG	0.9356	0.1253	0.0656
FKFK	0.9919	0.0954	0.2292
FLFL	0.8538	0.1989	0.0144
HTHT	0.8971	0.0345	0.0425
Instituut	0.7196	0.0345	0.0269
KKKK	0.6765	0.0381	0.0263
MJMJ	0.8910	0.0454	0.0395
MTMT	0.9832	0.0663	0.0809
OIOI	0.8695	0.0309	0.1061
SOSO	0.8631	0.0645	0.0813
USUS	0.4687	0.0145	0.0680

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
ARAR	0.3508	0.0457
BGBG	0.0198	0.1450
FKFK	0.0808	0.5256
FLFL	0.0001	0.0351
HTHT	0.0678	0.0038
Instituut	0.0244	0.0175
KKKK	0.0328	0.0000
MJMJ	0.0644	0.0007
MTMT	0.1269	0.0307
OIOI	0.0950	0.1165
SOSO	0.0874	0.0650
USUS	0.0495	0.0143

Indices of the Coordinates that Contribute Most to Inertia for the Row Points

	Dim1	Dim2	Best
ARAR	1	0	1
BGBG	0	2	2
FKFK	2	2	2
FLFL	0	0	2
HTHT	1	0	1
Instituut	0	0	1
KKKK	0	0	1
MJMJ	0	0	1
MTMT	1	0	1
OIOI	2	2	2
SOSO	1	1	1
USUS	0	0	1

Squared Cosines for the Row Points

	Dim1	Dim2
ARAR	0.8689	0.0729
BGBG	0.1640	0.7716
FKFK	0.1913	0.8006
FLFL	0.0050	0.8488
HTHT	0.8657	0.0314
Instituut	0.4931	0.2266
KKKK	0.6765	0.0000
MJMJ	0.8845	0.0065
MTMT	0.8507	0.1325
OIOI	0.4860	0.3835
SOSO	0.5838	0.2794
USUS	0.3954	0.0733

Column Coordinates

	Dim1	Dim2
30-39	-0.1659	-0.1312
40-49	-0.2440	-0.0851
50-59	0.2364	0.1826
kuni 29	0.5941	-0.3107
yle 60	-0.0252	0.4160

Summary Statistics for the Column Points

	Quality	Mass	Inertia
30-39	0.7682	0.2961	0.1286
40-49	0.7880	0.2525	0.1595
50-59	0.8730	0.2153	0.1640
kuni 29	0.9880	0.1063	0.3604
yle 60	0.8968	0.1299	0.1876

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
30-39	0.1120	0.1088
40-49	0.2064	0.0391
50-59	0.1652	0.1532
kuni 29	0.5152	0.2190
yle 60	0.0011	0.4800

Indices of the Coordinates that Contribute Most to Inertia for the Column Points

	Dim1	Dim2	Best
30-39	0	0	1
40-49	1	0	1
50-59	1	1	1
kuni 29	1	1	1
yle 60	0	2	2

Squared Cosines for the Column Points

	Dim1	Dim2
30-39	0.4728	0.2954
40-49	0.7025	0.0855
50-59	0.5467	0.3263
kuni 29	0.7758	0.2121
yle 60	0.0033	0.8935

Tabel 2. Korrespondentsanalüüsi tulemused tunnustele 'ametinimetus' ja 'vanusgrupp'.

The CORRESP Procedure
Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	19 38 57 76 95
0.44724	0.20002	220.221	93.06	93.06	**********
0.12087	0.01461	16.085	6.80	99.86	**
0.01731	0.00030	0.330	0.14	100.00	
Total	0.21493	236.637	100.00		

Degrees of Freedom = 12

Row Coordinates

	Dim1	Dim2
Assistent	-0.1023	0.1454
Dotsent	0.3605	0.1205
Profeesor	0.8283	-0.2106
Teadur	-0.4153	-0.0647

Summary Statistics for the Row Points

	Quality	Mass	Inertia
Assistent	0.9630	0.1526	0.0233
Dotsent	0.9979	0.2743	0.1848
Profeesor	0.9998	0.1244	0.4230
Teadur	0.9999	0.4487	0.3689

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
Assistent	0.0080	0.2209
Dotsent	0.1782	0.2727
Profeesor	0.4268	0.3778
Teadur	0.3869	0.1286

Indices of the Coordinates that Contribute Most to Inertia for the Row Points

	Dim1	Dim2	Best
Assistent	0	2	2
Dotsent	0	2	2

Indices of the Coordinates that Contribute Most to Inertia for the Row Points

	Dim1	Dim2	Best
Profeesor	1	1	1
Teadur	1	0	1

Squared Cosines for the Row Points

	Dim1	Dim2
Assistent	0.3188	0.6442
Dotsent	0.8976	0.1003
Profeesor	0.9391	0.0607
Teadur	0.9762	0.0237

Column Coordinates

	Dim1	Dim2
30-39	-0.3612	0.1115
40-49	0.0688	0.0709
50-59	0.4070	-0.0390
kuni 29	-0.7944	-0.2733
yle 60	0.6652	-0.1037

Summary Statistics for the Column Points

	Quality	Mass	Inertia
30-39	0.9985	0.2961	0.1972
40-49	0.9268	0.2525	0.0124
50-59	0.9991	0.2153	0.1676
kuni 29	0.9999	0.1063	0.3490
yle 60	0.9999	0.1299	0.2739

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
30-39	0.1932	0.2517
40-49	0.0060	0.0870
50-59	0.1782	0.0224
kuni 29	0.3353	0.5431
yle 60	0.2873	0.0957

Indices of the Coordinates that Contribute Most to Inertia for the Column Points

	Dim1	Dim2	Best
30-39	2	2	2
40-49	0	0	2
50-59	0	0	1
kuni 29	2	2	2
yle 60	1	1	1

Squared Cosines for the Column Points

	Dim1	Dim2
30-39	0.9117	0.0868
40-49	0.4494	0.4774
50-59	0.9900	0.0091
kuni 29	0.8942	0.1058
yle 60	0.9762	0.0237

Tabel 3. Korrespondentsanalüüsi tulemused tunnustele 'teaduskraad' ja 'teaduskond'.

The CORRESP Procedure
Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	13 26 39 52 65
0.39823	0.15858	174.601	65.63	65.63	**********
0.28817	0.08304	91.429	34.37	100.00	**********
Total	0.24163	266.030	100.00		

Degrees of Freedom = 22

Row Coordinates

	Dim1	Dim2
dok	0.0170	-0.2694
krd	0.8052	0.3349
mag	-0.4305	0.2940

Summary Statistics for the Row Points

	Quality	Mass	Inertia
dok	1.0000	0.5332	0.1607
krd	1.0000	0.1553	0.4888
mag	1.0000	0.3115	0.3504

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
dok	0.0010	0.4659
krd	0.6349	0.2098
mag	0.3641	0.3243

Indices of the Coordinates that Contribute Most to Inertia for the Row Points

	Dim1	Dim2	Best
dok	0	2	2
krd	1	1	1
mag	1	1	1

Squared Cosines for the Row Points

	Dim1	Dim2
dok	0.0040	0.9960
krd	0.8525	0.1475
mag	0.6819	0.3181

Column Coordinates

	Dim1	Dim2
ARAR	0.6689	-0.0442
BGBG	-0.3482	-0.2547
FKFK	-0.2034	-0.5065
FLFL	-0.1141	0.3499
HTHT	-0.2954	0.5282
Instituut	-0.0813	0.0100
KKKK	-0.1626	0.3457
MJMJ	-0.2485	0.0150
MTMT	-0.2381	-0.3704
OIOI	-0.1620	0.2362
SOSO	-0.3681	0.0921
USUS	-0.5192	0.0428

Summary Statistics for the Column Points

	Quality	Mass	Inertia
ARAR	1.0000	0.2516	0.4679
BGBG	1.0000	0.1253	0.0965
FKFK	1.0000	0.0954	0.1176
FLFL	1.0000	0.1989	0.1115
HTHT	1.0000	0.0345	0.0523
Instituut	1.0000	0.0345	0.0010
KKKK	1.0000	0.0381	0.0230
MJMJ	1.0000	0.0454	0.0116
MTMT	1.0000	0.0663	0.0532
OIOI	1.0000	0.0309	0.0105
SOSO	1.0000	0.0645	0.0384
USUS	1.0000	0.0145	0.0163

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
ARAR	0.7099	0.0059
BGBG	0.0958	0.0979
FKFK	0.0249	0.2946
FLFL	0.0163	0.2933
HTHT	0.0190	0.1160
Instituut	0.0014	0.0000
KKKK	0.0064	0.0549
MJMJ	0.0177	0.0001
MTMT	0.0237	0.1095
OIOI	0.0051	0.0207
SOSO	0.0551	0.0066
USUS	0.0247	0.0003

Indices of the Coordinates that Contribute Most to Inertia for the Column Points

	Dim1	Dim2	Best
ARAR	1	0	1
BGBG	2	0	2
FKFK	0	2	2
FLFL	0	2	2
HTHT	0	2	2
Instituut	0	0	1
KKKK	0	0	2
MJMJ	0	0	1
MTMT	0	2	2
OIOI	0	0	2
SOSO	0	0	1
USUS	0	0	1

Squared Cosines for the Column Points

	Dim1	Dim2
ARAR	0.9957	0.0043
BGBG	0.6513	0.3487
FKFK	0.1389	0.8611
FLFL	0.0961	0.9039
HTHT	0.2382	0.7618
Instituut	0.9851	0.0149
KKKK	0.1811	0.8189
MJMJ	0.9964	0.0036
MTMT	0.2923	0.7077
OIOI	0.3201	0.6799
SOSO	0.9411	0.0589
USUS	0.9933	0.0067

Tabel 4. Korrespondentsanalüüsi tulemused tunnustele 'teaduskraad' ja 'vanusgrupp'.

The CORRESP Procedure
Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	19 38 57 76 95
0.45063	0.20307	223.577	97.31	97.31	************************
0.07486	0.00560	6.170	2.69	100.00	*
Total	0.20867	229.747	100.00		

Degrees of Freedom = 8

Row Coordinates

	Dim1	Dim2
dok	-0.3862	-0.0281
krd	0.0659	0.1742
mag	0.6280	-0.0387

Summary Statistics for the Row Points

	Quality	Mass	Inertia
dok	1.0000	0.5332	0.3831
krd	1.0000	0.1553	0.0258
mag	1.0000	0.3115	0.5911

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
dok	0.3916	0.0753
krd	0.0033	0.8414
mag	0.6051	0.0833

Indices of the Coordinates that Contribute Most to Inertia for the Row Points

	Dim1	Dim2	Best
dok	1	0	1
krd	0	2	2
mag	1	0	1

Squared Cosines for the Row Points

	Dim1	Dim2
dok	0.9947	0.0053
krd	0.1251	0.8749
mag	0.9962	0.0038

Column Coordinates

	Dim1	Dim2
30-39	0.2582	0.0972
40-49	-0.1270	0.0062
50-59	-0.3322	-0.0353
kuni 29	1.0029	-0.1380
yle 60	-0.6118	-0.0623

Summary Statistics for the Column Points

	Quality	Mass	Inertia
30-39	1.0000	0.2961	0.1080
40-49	1.0000	0.2525	0.0196
50-59	1.0000	0.2153	0.1151
kuni 29	1.0000	0.1063	0.5219
yle 60	1.0000	0.1299	0.2354

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
30-39	0.0972	0.4993
40-49	0.0201	0.0017
50-59	0.1170	0.0478
kuni 29	0.5264	0.3611
yle 60	0.2394	0.0901

Indices of the Coordinates that Contribute Most to Inertia for the Column Points

	Dim1	Dim2	Best
30-39	0	2	2
40-49	0	0	1
50-59	1	0	1
kuni 29	1	1	1
yle 60	1	0	1

Squared Cosines for the Column Points

	Dim1	Dim2
30-39	0.8758	0.1242
40-49	0.9976	0.0024
50-59	0.9889	0.0111
kuni 29	0.9814	0.0186
yle 60	0.9897	0.0103

Tabel 5. Korrespondentsanalüüsi tulemused tunnustele 'teaduskraad' ja 'sünnikoht'.

The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	19 38 57 76 95
0.11014	0.01213	12.5543	94.35	94.35	**********
0.02694	0.00073	0.7514	5.65	100.00	*
Total	0.01286	13.3057	100.00		

Degrees of Freedom = 6

Row Coordinates

	Dim1	Dim2
dok	-0.0446	-0.0232
krd	0.2527	0.0022
mag	-0.0535	0.0374

Summary Statistics for the Row Points

	Quality	Mass	Inertia
dok	1.0000	0.5246	0.1031
krd	1.0000	0.1594	0.7921
mag	1.0000	0.3159	0.1048

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
dok	0.0859	0.3894
krd	0.8395	0.0011
mag	0.0746	0.6095

Indices of the Coordinates that Contribute Most to Inertia for the Row Points

	Dim1	Dim2	Best
dok	0	2	2
krd	1	0	1
mag	0	2	2

Squared Cosines for the Row Points

	Dim1	Dim2
dok	0.7867	0.2133
krd	0.9999	0.0001
mag	0.6715	0.3285

Column Coordinates

	Dim1	Dim2
Ida	0.1150	0.0199
Laas	-0.1413	0.0609
Louna	0.0661	-0.0063
Pohi	-0.1723	-0.0321

Summary Statistics for the Column Points

	Quality	Mass	Inertia
Ida	1.0000	0.1101	0.1167
Laas	1.0000	0.1246	0.2295
Louna	1.0000	0.5739	0.1966
Pohi	1.0000	0.1913	0.4572

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
Ida	0.1201	0.0600
Laas	0.2052	0.6361
Louna	0.2064	0.0316
Pohi	0.4683	0.2723

Indices of the Coordinates that Contribute Most to Inertia for the Column Points

	Dim1	Dim2	Best
Ida	0	0	1
Laas	2	2	2
Louna	1	0	1
Pohi	1	1	1

Squared Cosines for the Column Points

	Dim1	Dim2
Ida	0.9710	0.0290
Laas	0.8435	0.1565
Louna	0.9909	0.0091
Pohi	0.9664	0.0336

Tabel 6. Korrespondentsanalüüsi tulemused tunnustele 'teaduskond' ja 'sünnikoht'.

The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	14 28 42 56 70
0.15174	0.02302	23.8307	72.36	72.36	**************************
0.07633	0.00583	6.0300	18.31	90.67	*******
0.05447	0.00297	3.0710	9.33	100.00	***
Total	0.03182	32.9318	100.00		

Degrees of Freedom = 33

Row Coordinates

	Dim1	Dim2
ARAR	-0.1535	-0.0493
BGBG	0.1716	-0.0524
FKFK	-0.0109	0.0390
FLFL	0.0907	0.0392
HTHT	-0.1870	-0.0123
Instituut	0.2801	-0.1418
KKKK	-0.1742	0.0947
MJMJ	-0.2240	-0.0463
MTMT	-0.0550	0.0745
OIOI	0.1200	0.2165
SOSO	0.2341	-0.0652
USUS	0.0949	0.2736

Summary Statistics for the Row Points

	Quality	Mass	Inertia
ARAR	0.9270	0.2599	0.2289
BGBG	0.9868	0.1285	0.1318
FKFK	0.5343	0.0918	0.0089
FLFL	0.9866	0.1874	0.0583
HTHT	0.9580	0.0338	0.0390
Instituut	0.9549	0.0338	0.1097
KKKK	0.6858	0.0396	0.0713
MJMJ	0.7613	0.0464	0.1001
MTMT	0.7432	0.0696	0.0252
OIOI	0.9624	0.0329	0.0658
SOSO	0.9725	0.0618	0.1181
USUS	0.8889	0.0145	0.0430

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
ARAR	0.2658	0.1084
BGBG	0.1644	0.0605
FKFK	0.0005	0.0240
FLFL	0.0669	0.0494
HTHT	0.0513	0.0009
Instituut	0.1152	0.1167
KKKK	0.0522	0.0609
MJMJ	0.1010	0.0171
MTMT	0.0091	0.0664
OIOI	0.0206	0.2644
SOSO	0.1472	0.0452
USUS	0.0057	0.1862

Indices of the Coordinates that Contribute Most to Inertia for the Row Points

	Dim1	Dim2	Best
ARAR	1	1	1
BGBG	1	0	1
FKFK	0	0	2
FLFL	1	0	1
HTHT	0	0	1
Instituut	2	2	2
KKKK	0	2	2
MJMJ	1	0	1
MTMT	0	2	2
OIOI	0	2	2
SOSO	1	0	1
USUS	0	2	2

Squared Cosines for the Row Points

	Dim1	Dim2
ARAR	0.8403	0.0867
BGBG	0.9027	0.0841
FKFK	0.0385	0.4958
FLFL	0.8314	0.1551
HTHT	0.9539	0.0041
Instituut	0.7600	0.1949
KKKK	0.5294	0.1564
MJMJ	0.7301	0.0312
MTMT	0.2619	0.4813
OIOI	0.2262	0.7362
SOSO	0.9024	0.0701
USUS	0.0955	0.7934

Column Coordinates

	Dim1	Dim2
Ida	-0.0186	0.1971
Laas	0.1737	0.0543
Louna	-0.1159	-0.0300
Pohi	0.2453	-0.0589

Summary Statistics for the Column Points

	Quality	Mass	Inertia
Ida	0.9043	0.1101	0.1500
Laas	0.6820	0.1246	0.1903
Louna	0.9991	0.5739	0.2588
Pohi	0.9547	0.1913	0.4009

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
Ida	0.0017	0.7343
Laas	0.1633	0.0631
Louna	0.3350	0.0885
Pohi	0.5000	0.1141

Indices of the Coordinates that Contribute Most to Inertia for the Column Points

	Dim1	Dim2	Best
Ida	0	2	2
Laas	0	0	1
Louna	1	0	1
Pohi	1	1	1

Squared Cosines for the Column Points

	Dim1	Dim2
Ida	0.0080	0.8963
Laas	0.6212	0.0607
Louna	0.9365	0.0626
Pohi	0.9026	0.0521

Lisa 2 Programmid

Programm 1. Korrespondentsanalüüs tunnustele 'teaduskond' ja 'vanusgrupp'.

```
data proov;
set minu.semester;
if vanusko=1 then vanusko1="kuni 29";
if vanusko=2 then vanusko1="30-39";
if vanusko=3 then vanusko1="40-49";
if vanusko=4 then vanusko1="50-59";
if vanusko=5 then vanusko1="yle 60";
keep teadusko vanusko1;
proc corresp data=proov outc=joonis;
tables teadusko, vanusko1 ;
run;
%plotit (color=black, data=joonis, datatype=corresp)
```

Programm 2. Korrespondentsanalüüs tunnustele 'ametinimetus' ja 'vanusgrupp'.

```
data proov;
set minu.semester;
if ametko=1 then ametko1="Professor";
if ametko=2 then ametko1="Dotsent";
if ametko=3 then ametko1="Teadur";
if ametko=4 then ametko1="Assistent";
if vanusko=1 then vanusko1="kuni 29";
if vanusko=2 then vanusko1="30-39";
if vanusko=3 then vanusko1="40-49";
if vanusko=4 then vanusko1="50-59";
if vanusko=5 then vanusko1="yle 60";
keep ametko1 vanusko1;
proc corresp data=proov outc=joonis;
tables  ametko1, vanusko1 ;
run;
%plotit (color=black, data=joonis, datatype=corresp)
```

Programm 3. Korrespondentsanalüüs tunnustele 'teaduskraad' ja 'teaduskond'.

```
data proov;
set minu.semester;
if kraadko=1 then kraadko1="dok";
if kraadko=2 then kraadko1="mag";
if kraadko=3 then kraadko1="krd";
keep teadusko kraadko1;
proc corresp data=proov outc=joonis;
tables  kraadko1, teadusko ;
run;
%plotit (color=black, data=joonis, datatype=corresp)
```

Programm 4. Korrespondentsanalüüs tunnustele 'teaduskraad' ja 'vanusgrupp'.

```
data proov;
set minu.semester;
if kraadko=1 then kraadko1="dok";
if kraadko=2 then kraadko1="mag";
if kraadko=3 then kraadko1="krd";
if vanusko=1 then vanusko1="kuni 29";
if vanusko=2 then vanusko1="30-39";
if vanusko=3 then vanusko1="40-49";
if vanusko=4 then vanusko1="50-59";
if vanusko=5 then vanusko1="yle 60";
keep vanusko1 kraadko1;
proc corresp data=proov outc=joonis;
tables  kraadko1, vanusko1 ;
run;
%plotit (color=black, data=joonis, datatype=corresp)
```

Programm 5. Korrespondentsanalüüs tunnustele 'teaduskraad' ja 'sünnikoht'.

```
data abi;
set minu.semester;
if kraadko=1 then kraadko1="dok";
if kraadko=2 then kraadko1="mag";
if kraadko=3 then kraadko1="krd";
if synniko=1 then synniko1="Louna";
if synniko=10 then synniko1="Louna";
if synniko=13 then synniko1="Louna";
if synniko=14 then synniko1="Louna";
if synniko=15 then synniko1="Louna";
if synniko=16 then synniko1="Louna";
if synniko=17 then synniko1="Louna";
if synniko=2 then synniko1="Pohi";
if synniko=5 then synniko1="Pohi";
if synniko=9 then synniko1="Ida";
if synniko=6 then synniko1="Ida";
if synniko=7 then synniko1="Ida";
if synniko=3 then synniko1="Laas";
if synniko=4 then synniko1="Laas";
if synniko=8 then synniko1="Laas";
if synniko=11 then synniko1="Laas";
if synniko=12 then synniko1="Laas";
keep kraadko1 synniko1;
proc corresp data=abi outc=joonis;
tables  kraadko1, synniko1 ;
run;
%plotit (color=black, data=joonis, datatype=corresp)
```

Programm 6. Korrespondentsanalüüs tunnustele 'teaduskond' ja 'sünnikoht'.

```
data abi;
set minu.semester;
if synniko=1 then synniko1="Louna";
if synniko=10 then synniko1="Louna";
if synniko=13 then synniko1="Louna";
if synniko=14 then synniko1="Louna";
if synniko=15 then synniko1="Louna";
if synniko=16 then synniko1="Louna";
if synniko=17 then synniko1="Louna";
if synniko=2 then synniko1="Pohi";
if synniko=5 then synniko1="Pohi";
if synniko=9 then synniko1="Ida";
if synniko=6 then synniko1="Ida";
if synniko=7 then synniko1="Ida";
if synniko=3 then synniko1="Laas";
if synniko=4 then synniko1="Laas";
if synniko=8 then synniko1="Laas";
if synniko=11 then synniko1="Laas";
if synniko=12 then synniko1="Laas";
keep teadusko synniko1;
proc corresp data=abi outc=joonis;
tables  teadusko, synniko1 ;
run;
%plotit (color=black, data=joonis, datatype=corresp)
```