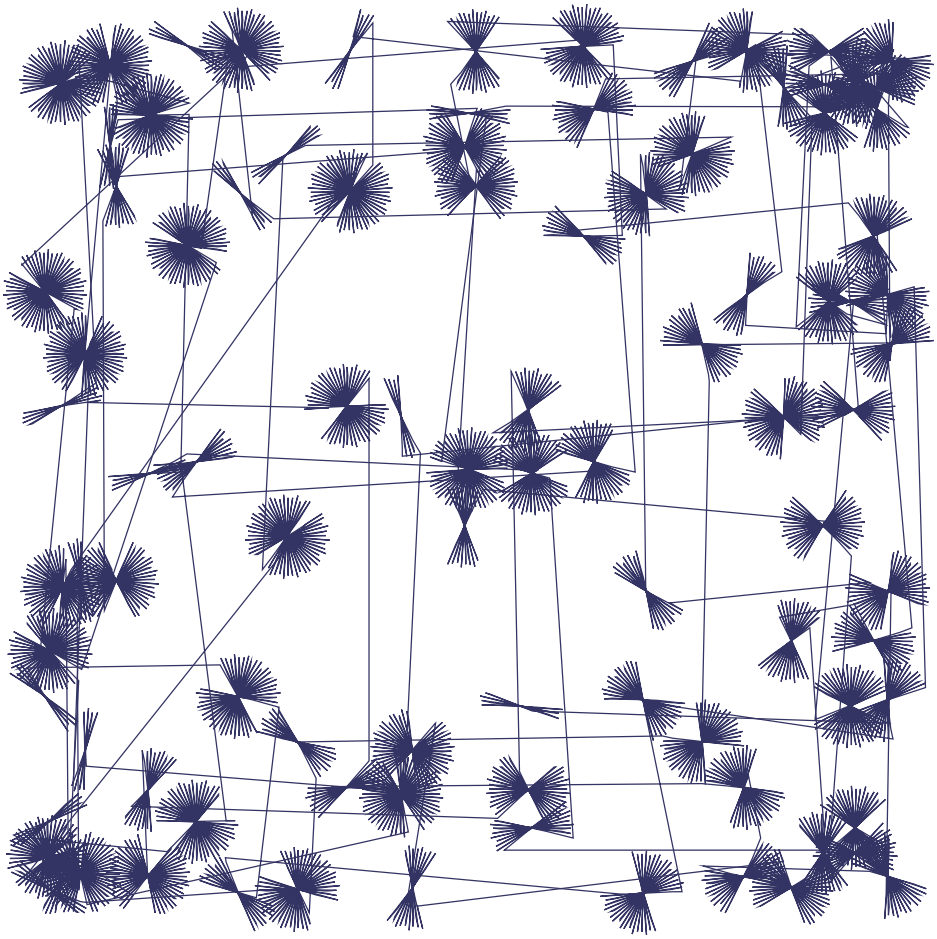
Tilastolliset monimuuttujamenetelmät

Seppo Mustonen
Helsingin yliopisto
Tilastotieteen laitos
1995



Tämä on verkkoversio kirjasta

Seppo Mustonen: Tilastolliset monimuuttujamenetelmät, Survo Systems Oy (1995).

Kirja on suunniteltu, kirjoitettu ja tulostettu PostScript-tiedostoiksi
Survo-ohjelmiston avulla.
Myös kaikki laskelmat ja analyysit on tehty Survolla.

PostScript-tiedostot on yhdistetty ja muunnettu PDF-tiedostoksi
www.survo.fi/mustonen/monim.pdf
Ghostscript- ja Adobe Acrobat- ohjelmilla.

Esipuhe

Tilastollisilla monimuuttujamenetelmillä käsitellään nimensä mukaisesti usean satunnaismuuttujan aineistoja. Koska muuttujia voi olla kymmeniä - jopa satoja, yleisenä pyrkimyksenä on vähentää muuttujien määrää tai yhdistellä muuttujia sopivien sääntöjen mukaan. Koko aineistoon liittyvästä vaihtelusta yritetään siis karsia puhtaasti satunnainen osuus tiivistämällä tietoa ja näin ehkä saadaan paljastetuksi tutkittavan ilmiön taustalla olevat rakenteet. Edellä sanottu koskee erilaisia kuvausmenetelmiä, joita ovat esim. pääkomponenttianalyysi, faktorianalyysi, kanoniset korrelaatiot, erotteluanalyysi ja ryhmittelyanalyysi.

Monimuuttujamenetelmien piiriin voi lukea myös suorat yhden muuttujan menetelmien yleistykset. Näin on mm. eräiden keskeisten tilastollisten testien laita. Esim. tavallinen *t*-testi yleistyy usean muuttujan tapauksessa Hotellingin T^2-testiksi.

Monimuuttujamenetelmäksi ei sen sijaan katsota esim. usean selittävän muuttujan regressioanalyysia, koska tässä tapauksessa satunnaisena muuttujana käsitellään vain selitettävää muuttujaa; selittäjät voivat olla esim. koesuunnittelun määräämiä systemaattisia tekijöitä. Luonnollisesti regressioanalyysia sovelletaan kuitenkin usein rinnan monimuuttuja-analyysien kanssa. Tyypillinen toimintatapa saattaa olla se, että aluksi jollakin monimuuttujamenetelmällä "puhdistetaan" selittävien muuttujien joukkoa vähentämällä muuttujien määrää ja/tai tekemällä ne vähemmän toisistaan riippuviksi. Lopullinen tarkastelu tapahtuu regressioanalyysilla tämän puhdistuksen jälkeen.

Monimuuttujamenetelmien suosio on vaihdellut niiden koko olemassaoloajan aina 1930-luvulta lähtien. Laskennallisten hankaluuksien vuoksi soveltaminen tositilanteissa saattoi alkaa vasta 1950-luvulla tietokoneiden ansiosta. Tästä seurannut käytön helpottuminen ja eräiden menetelmien, ennen muuta faktorianalyysin, houkuttelevuus mekaanisiin soveltamisyrityksiin johti Suomessakin 1960-luvulla etenkin yhteiskunta- ja käyttäytymistieteiden piirissä laajamittaiseen ja joskus varsin perustelemattomaan käyttöön. Tämän ylimitoitetun suosion romahdukseen vaikutti osaltaan 1960-1970-lukujen vaihteessa vallinnut "positivistisen tieteen kritiikki". Tuntuu siltä, että olisi taas aika tuon menetetyn maineen palautua kohtuullisiin mittoihin.

Monimuuttujamenetelmien niin kuin monien muidenkin teknisesti vaativien tilastollisten keinojen opettamisen ongelmana on se, että, ne jotka näitä menetelmiä tarvitsevat, eivät yleensä pysty kunnolla omaksumaan menetelmien taustalla olevaa matemaattispitoista teoriaa. Tämän taustan ymmärtäminen on tärkeää ainakin siltä osin, mikä liittyy menetelmien käytön ehtoihin ja rajoituksiin. Sen sijaan esim. joidenkin otos- tai testisuureiden jakaumien johtaminen, mikä vaatii enemmän matemaattisen analyysin tuntemista, ei ole yhtä tärkeää, koska tällaiset tulokset on hyödynnettävissä ilman, että osaisi ne itse päätellä.

Useat etupäässä tilastotieteen puolella kirjoitetut oppikirjat ovat turhan teknisiä ja vailla yhteyksiä todellisiin sovelluksiin. Soveltajille tarkoitetuissa esityksissä taas teoreettinen puoli saatetaan jopa sivuuttaa tai tarjota enemmälti ilman perusteluja.

Omaksumassani lähestymistavassa olen pyrkinyt välttämään liikaa matematiikkaa. Opettaessani aihetta yli 30 vuoden ajan olen päätynyt ratkaisuun, jossa perusteoria eli multinormaalijakauman ominaisuudet saatetaan johtaa hyvin vähin eväin lähtemällä liikkeelle ko. jakauman konstruktiivisesta määritelmästä. Tällöin matematiikan osalta tärkeimmäksi perusvaatimukseksi nousee vain matriisilaskennan hallinta. Erityisen suuri merkitys on matriisin singulaariarvohajotelmalla ja eräillä siihen liittyvillä tuloksilla, joiden avulla useat monimuuttujamallit ovat helpoiten johdettavissa. Tätä keinoa ei jostain syystä ole mainittavasti käytetty alan oppikirjoissa. Vaikka matriisilaskennan merkintöjä ja perusteita ei tässä yhteydessä kerrata, liitteessä 2 on käyty läpi tarvittavat singulaariarvohajotelmaan liittyvät tarkastelut.

Asioiden geometrinen hahmottaminen on monille tärkeää. On mielenkiintoista mutta samalla valitettavaa, että kykymme osoittautuvat vajavaisiksi yleistäessämme 2- tai 3-ulotteisia mielikuviamme useampiulotteisiin avaruuksiin, joissa monimuuttujamenetelmiä koskevat tarkastelut yleensä liikkuvat. Esim. on vaikea tajuta intuitiivisesti jo sitä, että yleisissä peräkkäisissä koordinaatiston kierroissa vain kaksiulotteisessa tapauksessa ei ole väliä sillä missä järjestyksessä kierrot tehdään. Liite 1 sisältää esimerkkejä moniulotteisuuteen liittyvistä ongelmista ja samalla näytteitä siitä miten geometrinen ajattelutapa korvataan analyyttisella.

Tämä esitys tukeutuu monella tavalla ATK-tekniikan suomiin mahdollisuuksiin. Kaikki numeeriset esimerkit on kuvattu Survo-järjestelmän avulla valmiina kaavioina. Itse asiassa on kysymys hypertekstistä, jonka eräs esitysmuoto on tässä paperilla. Koko esimerkkimateriaali on koottu levykkeelle, jolloin esimerkit, sovellukset ja simulointikokeet on toistettavissa Survon avulla joutumatta uudelleen kirjoittamaan ja kopioimaan aineistoa käsin. Liitteessä 3 kerrotaan esimerkkilevykkeestä hieman lisää.

Myös varsinainen tekstiosa kaavoineen ja kuvineen on laadittu Survolla ja on jatkuvasti hallittavissa. Esim. pitäessäni kurssia voin helposti poimia minkä tahansa osan tästä aineistosta ja heijastaa sen luokassa kankaalle täsmälleen samassa ulkoasussa kuin se esiintyy tekstissä. Siis valmiiksi piirrettyjä kalvoja ei tarvita lainkaan, vaan ne syntyvät opetuksen aikana. Samalla tavalla otetaan käyttöön esimerkit, toistetaan niitä koskevat analyysit ja muunnellaan sekä aineistoja että analysointitapoja jopa hetken mielijohteesta.

Hyvin paljon painoa on annettu simulointikokeille. Näissä kokeissa luodaan tiettyä monimuuttujamallia vastaavia satunnaisotoksia, joita sitten analysoidaan vastaavilla malleilla. Tarkoituksena on näyttää, miten menetelmät toimivat eri tilanteissa. Todellisilla aineistoilla on paljon hankalampi selvittää menetelmien kyvykkyyttä, koska tuloksen ollessa huono on mahdotonta tietää, liittyykö ongelma itse menetelmään vai sopimattomaan aineistoon.

Nykyiset PC-laitteistot ovat jo niin tehokkaita, että esim. Fisherin satunnaistamisperiaatetta, jota yleensä on käytetty pienten yhden muuttujan aineistojen tutkimisessa, voidaan soveltaa myös todellisiin usean muuttujan otoksiin ja johtaa aineistokohtaisia testisuureiden kriittisiä rajoja simuloimalla. Näin on tehty tässä esityksessä faktorianalyysiin liittyvän tulosten vertailumenetelmän, transformaatioanalyysin residuaalien tarkastelussa.

Vaikka esitystapa edellä kuvatussa suhteessa on hyvin Survo-painotteinen, sen ei pitäisi olla esteenä tämän tekstin käytölle tavanomaisena oppikirjana, koska esimerkkikaaviot joko selittävät itsensä tai ne on varustettu asianmukaisilla kommenteilla. Survoa hallitseva lukija kuitenkin hyötyy lisää em. hypertekstiominaisuuksista.

Monimuuttujamenetelmiin liittyvää tietoa on nykyisin suunnattomasti. On pakko rajoittua joihinkin keskeisiksi koettuihin asioihin. Tämän esityksen painopiste on ehdottomasti klassisissa perusmenetelmissä, joiden tunteminen kuuluu tilastotieteilijän yleissivistykseen.

Aloitamme konkreettisesta päästä eli esittelemme keinoja, joilla moniulotteisia aineistoja kuvataan graafisesti. Sen jälkeen siirrytään suoraan multinormaalijakauman määritelmään ja sen perusominaisuuksiin. Tämä toimii välttämättömänä taustana itse menetelmien johtamiselle. Oman lukunsa muodostaa multinormaalijakauman otossuureiden, keskiarvojen, varianssien ja kovarianssien tarkastelu ja jakaumaan liittyvät testit. Lopuksi käydään läpi varsinaiset monimuuttujamallit.

Olisi suotavaa, että lukija heti alusta pitäen myös omilla aineistoillaan kokeilisi esiteltyjä keinoja. Tähän esitykseen on ollut mahdoton liittää kovin monta todellista tutkimustilannetta, koska jokaisen sellaisen pohjustaminen kulloistakin sovellusalaa tuntemattomille veisi kohtuuttomasti tilaa.

Tämä teksti pohjautuu puolittain aikaisempina vuosina pitämiini kursseihin. Nykyinen sisällys muotoutui kevätlukukauden 1994 luentojen aikana ja olen sitä vielä jonkin verran laajentanut kesällä.

Kiitän erityisesti *Seppo Hassia* ja *Simo Puntasta* monista arvokkaista parannuksista varsinkin liitteen 2 osalta. Samoin esitän kiitokseni *Jevgeni Koeville*, *Anna-Riitta Niskaselle*, *Marco Varjukselle* ja *Kimmo Vehkalahdelle* kevään 1994 kurssin aikana ja sen jälkeen saamastani palautteesta.

Hituniemessä joulukuussa 1994

S.M.

Sisällysluettelo

1. Kuvallisia keinoja

Tilastollisen aineiston graafisen esittämisen ongelmat korostuvat moniulotteisissa aineistoissa, sillä esim. monikymmenulotteisen pisteparven litistäminen tasoon tarkkuudesta tinkimättä on täysi mahdottomuus. Kolmiulotteisuus esim. stereokuvapareina tai kuvaruudulla pyörivinä ns. spin-kuvina ei tuota juuri mitään lisähyötyä näissä tilanteissa. Parasta on tunnustaa tosiasiat ja esittää se, mikä esitettävissä on, tasossa.

Kuten tulemme näkemään, eräät menetelmät tuottavat monen muuttujan aineistoista vähäulotteisia esityksiä esim. karsimalla tutkittavan ilmiön kannalta tarpeetonta satunnaisuutta. Tällöin menetelmien tuloksia tarkasteltaessa graafiset keinot tulevat paremmin ulottuvillemme.

Sopii kysyä, onko moniulotteisen ilmiön graafisessa esittämisessä mitään mieltä, koska itse ilmiöllä on harvoin suoraa suhdetta fysikaaliseen, näkyvään todellisuuteen. Kaikki kuvalliset keinot ovat tällöin täysin sopimuksenvaraisia. On kuitenkin kiistatonta, että ihmisen on jopa huonostikin suunnitellusta kuvallisesta esityksestä helpompi nähdä asioiden välisiä yhteyksiä kuin katselemalla pelkkää lukujen muodostamaa havaintomatriisia. Kuvien hahmottamisessa ihminen on jatkuvasti ylivoimainen tehokkaimpiinkin tietokoneratkaisuihin verrattuna.

Miltei kaikkien kuvallisten keinojen perustana ovat tavanomaiset kaksiulotteiset, suorakulmaiset koordinaattiesitykset, joissa havainnot näkyvät pisteinä tai pisteen laajennuksina. Laajennuksella tarkoitetaan sitä, että "pisteet" voivat olla erikokoisia, -muotoisia ja -värisiä. Niiden ympärille voi kasautua myös eri muuttujista riippuvaa tietoa erimittaisilla ja -suuntaisilla janoilla tai käyränpätkillä kuvattuina. Siis erilaisilla pisteen liitännäisillä saadaan kuvaan jollain tavoin mukaan hyvinkin monen muuttujan osuus.

On voitu päätyä hyvinkin erikoistuneisiin ratkaisuihin, joista eksoottisimpia ovat ns. Chernoffin naamat. Niissä muuttujat asetetaan vastaamaan kasvon eri piirteitä. Menetelmän viehätys piilee siinä, että tukeudutaan suoraan ihmisen opittuun kykyyn tunnistaa lähimmäisensä kasvoista.

1.1 Hajontakuvien yleistykset

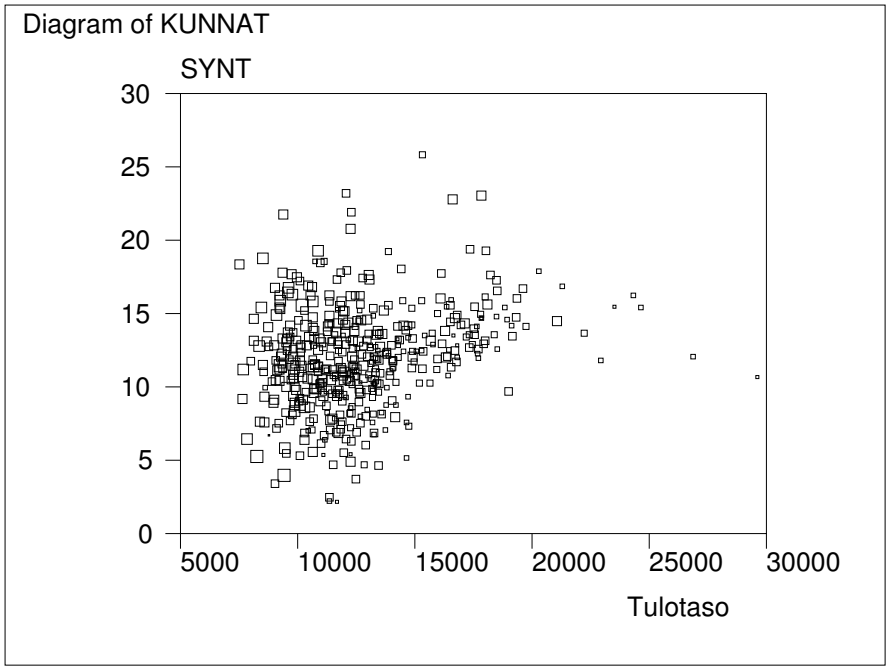
Kahden muuttujan hajontakuvissa, joita myös kutsutaan korrelaatiodiagrammoiksi, tarkastellaan ko. muuttujien keskinäisiä riippuvuuksia. Kutakin havaintoa vastaa kaksiulotteisessa koordinaatistossa piste, jonka asema x-akselin suunnassa määräytyy ensimmäisen muuttujan arvon ja y-akselin suunnassa toisen muuttujan arvon mukaan.

Tähän kuvaustapaan voi lisätä tietoa muista muuttujista laajentamalla eri tavoin "pisteen" ulkoista muotoa. Survon grafiikassa tämä käy helpoiten käyttämällä `POINT`-täsmennystä yleisimmässä muodossaan, kuten tapahtuu seu-

raavassa esimerkissä. Tällöin otetaan mukaan kolmas muuttuja, joka vaikuttaa pistettä vastaavan symbolin kokoon. Myös LINE-täsmennys, varsinkin LINE=6 (kts. Survo-kirjan ss. 262-3), eri laajennuksineen tarjoaa vielä monipuolisempia ja useammasta muuttujasta riippuvia tehostuksia.

Seuraavassa diagrammassa on piirretty vastakkain Suomen kunnista (Survon esimerkkiaineisto KUNNAT) muuttujat Tulotaso ja SYNT (syntyneisyys 1000 asukasta kohti) siten, että kuntaa vastaavan neliömäisen "pisteen" sivun pituus on verrannollinen muuttujaan Äyriero=Äyri-12. Ko. muunnos veroäyri-muuttujassa tehdään, jotta ko. erot todella näkyisivät kuvassa.

```
 27  1 SURVO 84C EDITOR Thu Jul 21 17:53:39 1994          C:\M\MON\ 100 100 0
  1 *
  2 *VAR SYNT=1000*Synt./Väestö TO KUNNAT
  3 *VAR Äyriero=Äyri-12 TO KUNNAT
  4 *.....................................................................
  5 *GPLOT KUNNAT,Tulotaso,SYNT_
  6 *POINT=5,10,Äyriero,8
  7 *
```



Kuvasta ilmenee paitsi tulotason ja syntyneisyyden riippuvuus myös se, että veroäyri on odotetusti alhaisen tulotason kunnissa suurimmillaan ja korkean tulotason kunnissa pienimmillään.

Rivin 6 POINT-täsmennys määrää, mitä kussakin tapauksessa tulee pisteen paikalle. Ensimmäinen parametri 5 valitsee symboliksi avoimen neliön. Toinen parametri 10 ilmoittaa neliön peruskoon ja kolmas parametri (Äyriero) kokoon vaikuttavan muuttujan. Peruskokoa käytetään, kun Äyriero on viimeisen parametrin (8) suuruinen. Yleisesti neliön koko (sivun pituus) on suhteessa muuttujan Äyriero arvoon.

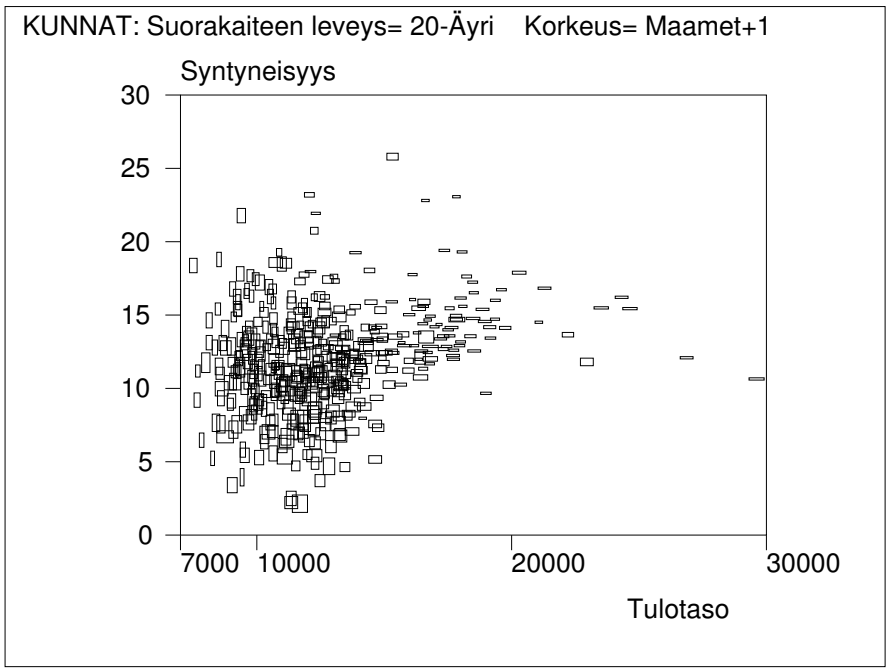
Neliön asemasta voidaan symboliksi valita suorakaide, jonka leveyttä ja korkeutta säädellään eri muuttujilla. Näin yhdessä kuvassa esitetään 4 eri muuttujan riippuvuutta samanaikaisesti.

Survossa tällaisen kuvan laatiminen edellyttää yleisempää piirrostekniikkaa, jossa piirrettävä symboli määritellään erillisten, muuttujan arvoista riippuvien janojen yhdistelmänä.

Piirroskaaviossa on erikseen annettu yleiset säädöt (rivit 13-19). Näillä kuvataan sellainen suorakaide, jonka keskipiste on (VX,VY), leveys vx ja korkeus vy. Suorakaide piirretään neljästä palasta (suorakaiteen sivut) koostuvana jatkuvana käyränä. Paloittelua säätelee parametri T, joka saa peräkkäin arvot 0,1,2,3,4 . Nämä yleiset säädöt muodostavat "piirrosohjelman", jota soveltajan ei edes tarvitse yksityiskohtaisesti ymmärtää.

Erikseen annetaan aineistokohtaiset säädöt (rivit 2-11), joilla soveltaja määrittelee aineistonsa (tässä DATA:KUNNAT) neljän muuttujan ja piirrosparametrien VX,VY,vx,vy vastaavuudet. Näiden joukkoon voi kuulua piirrosta yleisesti ohjaavia täsmennyksiä kuten tässä kuvan otsikko (HEADER rivillä 3), asteikot (XSCALE ja YSCALE rivillä 6) sekä akselien nimet (XLABEL ja YLABEL rivillä 7).

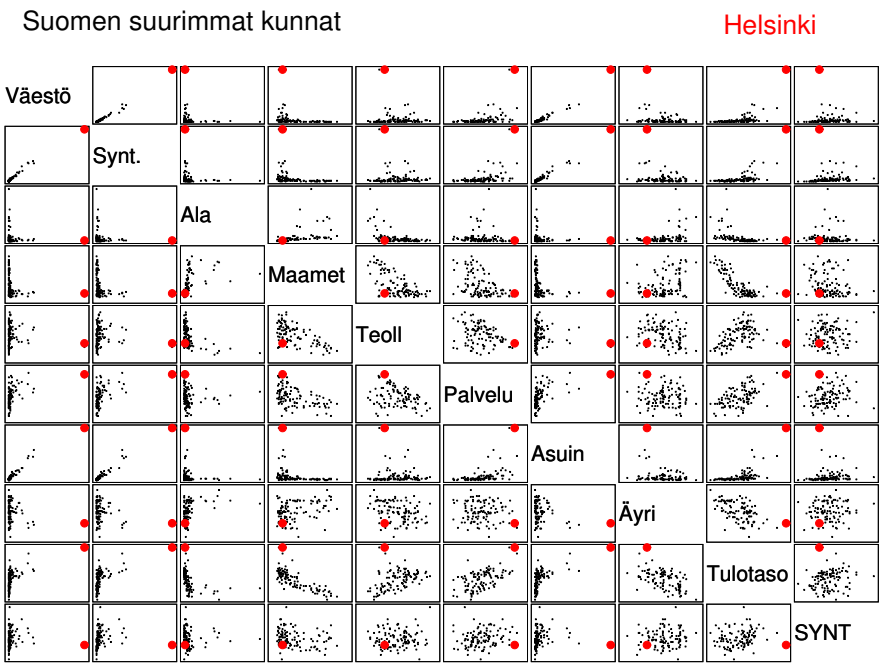
```
 22  1 SURVO 84C EDITOR Sat Jul 23 09:59:09 1994          C:\M\MON\ 100 100 0
  1 *
  2 *AINEISTOSTA RIIPPUVAT SÄÄDÖT:
  3 *  HEADER=KUNNAT:_Suorakaiteen_leveys=_20-Äyri____Korkeus=_Maamet+1
  4 *  X-Y-muuttujat:
  5 *     VX=DATA:KUNNAT,Tulotaso VY=DATA:KUNNAT,SYNT
  6 *     XSCALE=7000,10000,20000,30000 YSCALE=0(5)30
  7 *     XLABEL=Tulotaso               YLABEL=Syntyneisyys
  8 *  Suorakaidemuuttujat:
  9 *     Vx=DATA:KUNNAT,Äyri   Vy=DATA:KUNNAT,Maamet
 10 *     Muunnokset:
 11 *     vx=100*(20-Vx)    vy=(Vy+1)/6
 12 *
 13 *YLEISET SÄÄDÖT:  suorakaiteen leveys vx, korkeus vy, keskipiste (VX,VY)
 14 *  xx=VX-vx/2 yy=VY-vy/2 vasen alakulma
 15 *  T=0,4,1
 16 *  X1=if(T<=1)then(xx+T*vx)else(X2)  Y1=if(T<=1)then(yy)else(Y2)
 17 *  X2=if(T<=2)then(xx+vx)else(X3)    Y2=if(T<=2)then(yy+(T-1)*vy)else(Y3)
 18 *  X3=if(T<=3)then(xx+vx-(T-2)*vx)else(X4) Y3=if(T<=3)then(yy+vy)else(Y4)
 19 *  X4=xx                             Y4=yy+vy-(T-3)*vy
 20 *
 21 *GPLOT X(T)=X1,Y(T)=Y1_
 22 *
```



1.2 Hajontakuvamatriisit

Hajontakuvamatriisilla (Draftsman's display) tarkoitetaan kuvakoostetta, joka asettelultaan vastaa esim. korrelaatiomatriisia, mutta jonka "alkioina" ovat asianomaisten muuttujien korrelaatiodiagrammat. Englanninkielinen nimitys juontaa alkunsa teknisten laitteiden projektiopiirrostekniikasta. Tutkittavan aineiston kaikkien mahdollisten kaksiulotteisten hajontakuvien samanaikainen esittäminen antaa melko hyvän kokonaisnäkemyksen riippuvuuksien luonteesta. Se ei kuitenkaan voi tuottaa täydellistä kuvaa aineiston kokonaisvaihtelusta, koska minkäänulotteiset reunajakaumat eivät määrittele yhteisjakaumaa yksikäsitteisesti. Tästä huolimatta hajontakuvamatriisin piirtäminen on oivallinen keino tutustua uuden aineiston käyttäytymiseen ja auttaa esim. sopivien muuttujatransformaatioiden löytämisessä.

Suomen suurimpien kuntien (asukasluku yli 10000) 10 valitusta muuttujasta tehty hajontakuvamatriisi näyttää seuraavalta. Kussakin korrelaatiodiagrammassa Helsinki erottuu suurempana pisteenä.



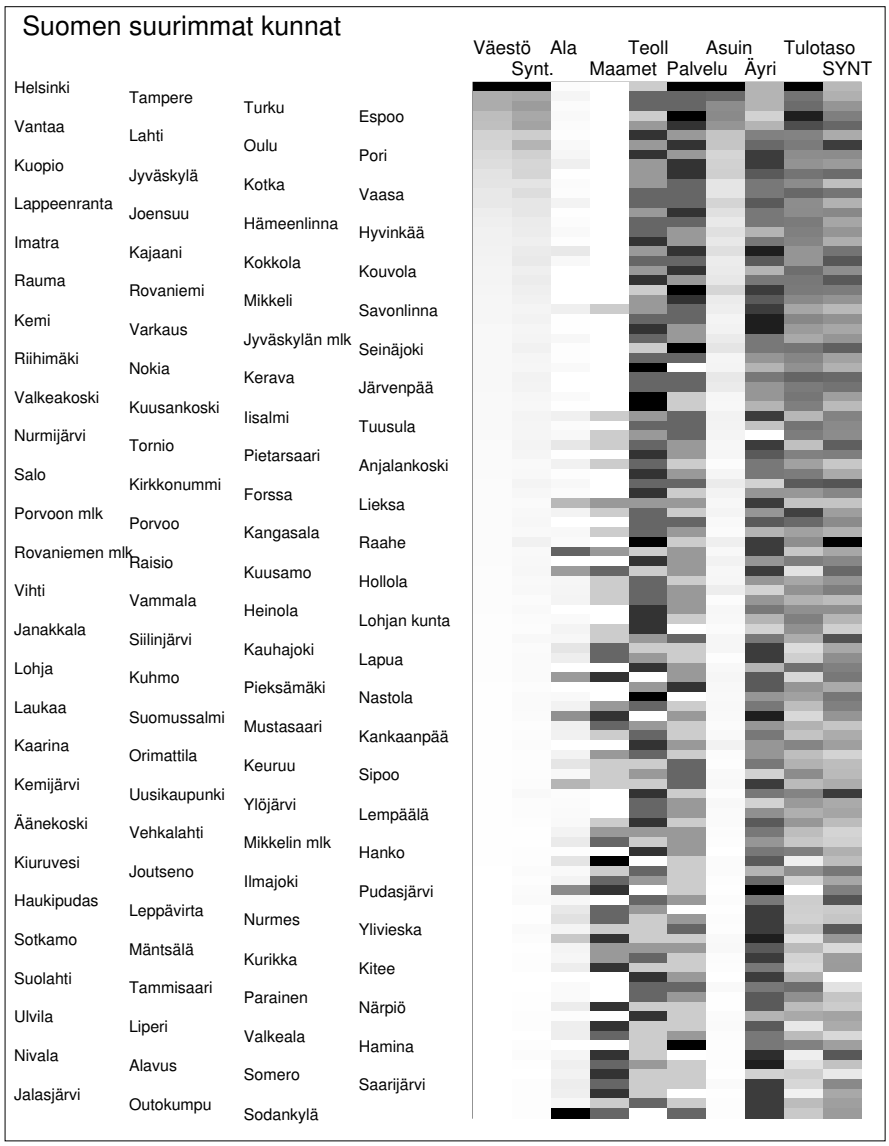
Tällainen kuva syntyy Survon avulla vähimmillään PLOT- (kuvaruutuun GPLOT-) komennolla, joka on varustettu täsmennyksellä TYPE=DRAFTS. Hyvin vähäluokkaisten muuttujien osalta on hyötyä JITTER-täsmennyksestä, joka täristää muuten päällekkäin tulevat pisteet "oikean" paikan ympärille satunnaiseksi pisteparveksi. Ilman täristystä diskreettien muuttujien hajontakuvat surkastuvat usein mielenkiinnottomiksi hilapisteistöiksi eikä riippuvuuden luonteesta saa kunnon käsitystä.

Tässä tapauksessa, koska yksi havainto halutaan erottaa muiden joukosta ja seurata sen asemaa kussakin osakuvassa erikseen, kuva rakennetaan kahdessa vaiheessa. Ensin piirretään koko aineisto (rivit 2-6) tallettaen sekä kuva (OUTFILE-täsmennys) että automaattisesti valitut piirrosasteikot (OUTSCALE-täsmennys). Tämän päälle saadaan sopeutettu kuva toisesta aineistosta (tai kuten tässä yhdestä havainnosta) eri värillä tai toisentyyppisinä pisteinä merkittynä muuten vastaavalla kaaviolla (rivit 8-14) käyttäen kohdistukseen INFILE- ja INSCALE-täsmennyksiä:

```
 13  1 SURVO 84C EDITOR Sat Jul 23 15:45:38 1994          C:\M\MON\ 100 100 0
  1 *
  2 *Koko aineiston piirto (skaalausten valinta ja täristys):
  3 *GPLOT KUNNAT  / TYPE=DRAFTS OUTSCALE=SKAALAT.TXT JITTER=30
  4 *IND=Väestö,10000,500000
  5 *XDIV=0,1,0 YDIV=0,10,1 HEADER=Suomen_suurimmat_kunnat
  6 *MASK=--AAAAAAAAAA MODE=VGA OUTFILE=A
  7 *.....................................................................
  8 *Yhden havainnon lisäys isommalla merkinnällä:
  9 *GPLOT KUNNAT_ / TYPE=DRAFTS INSCALE=SKAALAT.TXT
 10 *
 11 *XDIV=0,1,0 YDIV=0,10,1 HEADER=
 12 *MASK=--AAAAAAAAAA MODE=VGA INFILE=A POINT=[RED],0,3 TEXTS=Kunta
 13 *CASES=Kunta:Helsinki       Helsinki voidaan vaihtaa
 14 *Kunta=Helsinki,500,450     mihin tahansa muuhun kuntaan.
 15 *
```

1.3 Havaintomatriisi rasterikuvana

Toinen tapa yleiskuvan saamiseksi on piirtää koko havaintomatriisi matriisidiagrammana siten, että havaintoarvojen paikalla ovat niiden suuruuksia vastaavat tummuusasteeltaan vaihtelevat viivat tai laatikot. Esim. muuttujakohtaisesti säädetään, miten tummuusaste muuttuu mustasta valkoiseen arvon kasvaessa tai päinvastoin.



Esimerkkinämme on jälleen Suomen suurimpia kuntia kuvaavat 10 muuttujaa. Kunnat on järjestetty asukasluvun mukaan suurimmasta pienimpään, jolloin kuvasta voi päätellä helpommin, mitkä tiedot korreloivat hyvin asukasluvun

kanssa. Tässä esitystavassa kannattaa kiinnittää huomiota poikkeaviin tapauksiin, jotka näkyvät silmiinpistävinä "spektriviivoina".

Kuva on saatu aikaan seuraavasti:

```
 13  1 SURVO 84C EDITOR Sat Jul 23 17:40:34 1994          C:\M\MON\ 100 100 0
 15 *.....................................................................
 16 *IND=Väestö,10000,500000
 17 *FILE SORT KUNNAT BY -Väestö TO KUNNAT2
 18 *.....................................................................
 19 *MASK=A-AAAAAAAAAA--
 20 *HEADER=Suomen_suurimmat_kunnat
 21 *PLOT KUNNAT2_ /  TYPE=MATRIX SCREEN=NEG DEVICE=PS,KUNNAT4.PS
 22 *SIZE=1164,1500 XDIV=620,514,30 YDIV=30,1370,100
 23 *ROWLABELS=[Swiss(6)],1,4,10 COLUMNLABELS=[Swiss(7)],1,2
 24 *
```

Tiedoston KUNNAT mukaan otettavat havainnot on ensin lajiteltu väkiluvun mukaan laskevaan järjestykseen havaintotiedostoksi KUNNAT2 (rivit 16-17). Kuvan piirto tapahtuu riveillä 19-23 olevalla PLOT-kaaviolla, jossa kuvatyypin määrää täsmennys TYPE=MATRIX. Täsmennys SCREEN=NEG tarkoittaa, että muuttujanarvon vähetessä myös tummuusaste vähenee. Täsmennysten ROWLABELS ja COLUMNLABELS avulla rivi- ja sarakeotsikot saadaan lomittumaan niin, etteivät ne ahtaudu päällekkäin.

1.4 Andrews-käyrät

Kokonaan toisenlaisen näkökulman moniulotteisen aineiston graafiseen tarkasteluun tarjoaa *D.F.Andrews*in (1972) esittämä Fourier-käyrätekniikka. Kutakin p muuttujan $\mathbf{X}=(X_1,X_2,...,X_p)$ havaintoa vastaa funktion

$$f_{\mathbf{X}}(t) = X_1/\sqrt{2} + X_2\sin(t) + X_3\cos(t) + X_4\sin(2t) + X_5\cos(2t) + X_6\sin(3t) + ...$$

kuvaaja välillä $-\pi < t < \pi$. Kun havainnot esitetään samassa koordinaatistossa, toisiaan muistuttavia havaintoja edustavat luonnollisesti toisiaan muistuttavat käyrät. Käyrien etäisyys toisistaan vastaa jopa tarkkaan havaintojen euklidista etäisyyttä p-ulotteisessa avaruudessa siinä mielessä, että havainnoille $\mathbf{X}$ ja $\mathbf{Y}$ pätee

$$\frac{1}{\pi}\int_{-\pi}^{\pi}[f_{\mathbf{X}}(t) - f_{\mathbf{Y}}(t)]^2 dt = \|\mathbf{X}-\mathbf{Y}\|^2 = (X_1-Y_1)^2 + (X_2-Y_2)^2 + ... + (X_p-Y_p)^2 .$$

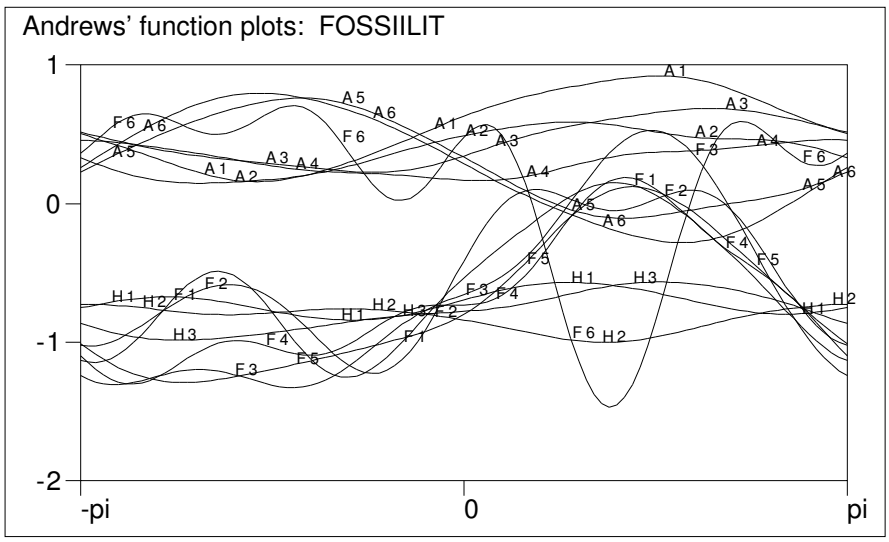
Andrews käytti eräänä esimerkkinään pientä ihmis- ja apinalajien sekä muinaisisten fossiilien leukaluista tehdyistä mittauksista koottua aineistoa. Alkuperäiset 8 muuttujaa on seuraavassa havaintotaulukossa korvattu erotteluanalyysin antamilla erottelumuuttujilla, jolloin eri lajien ja rotujen poikkeamat näkyvät muuttujissa X_1-X_8 voimakkuusjärjestyksessä. Yleensäkin muuttujat kannattaa asettaa tärkeysjärjestykseen, koska niiden vaikutukset itse käyrissä ilmenevät sitä paremmin. mitä alhaisemmasta "frekvenssistä" on kysymys.
Erityisesti ensimmäinen muuttuja (X_1) määrää yksinkertaisesti, millä perus-

tasolla havaintoa vastaava käyrä kulkee.

Aineiston kolme ensimmäistä havaintoa vastaavat nykyisiä ihmisrotuja (länsi-afrikkalainen, britti, australialainen), 6 seuraavaa tunnettuja apinalajeja ja loput 6 muinaisia löydöksiä. Mielenkiintoista on tarkastella viimeistä (Proconsul Africanus), jota ainakin joskus on pidetty apinoiden ja ihmisten välisenä "puuttuvana renkaana".

```
  1  1 SURVO 84C EDITOR Sun Jul 24 14:24:08 1994          C:\M\MON\ 150 100 0
  1 *
  2 *DATA FOSSIILIT
  3 * Laji       X1    X2    X3    X4    X5    X6    X7    X8     Tunnus
  4 * Westafr  -8.09  0.49  0.18  0.75 -0.06 -0.04  0.04  0.03    H1
  5 * British  -9.37 -0.68 -0.44 -0.37  0.37  0.02 -0.01  0.05    H2
  6 * Austral  -8.87  1.44  0.36 -0.34 -0.29 -0.02 -0.01 -0.05    H3
  7 * Gorilla1  6.28  2.89  0.43 -0.03  0.10 -0.14  0.07  0.08    A1
  8 * Gorilla2  4.28  1.52  0.71 -0.06  0.25  0.15 -0.07 -0.10    A2
  9 * Orang1    5.11  1.61 -0.72  0.04 -0.17  0.13  0.03  0.05    A3
 10 * Orang2    3.60  0.28 -1.05  0.01 -0.03 -0.11 -0.11 -0.08    A4
 11 * Chimpan1  3.46 -3.37  0.33 -0.32 -0.19 -0.04  0.09  0.09    A5
 12 * Chimpan2  3.05 -4.21  0.17  0.28  0.04  0.02 -0.06 -0.06    A6
 13 * Pith.Pek -6.73  3.63  1.14  2.11 -1.90  0.24  1.23 -0.55    F1
 14 * Pith.P2  -5.90  3.95  0.89  1.58 -1.56  1.10  1.53  0.58    F2
 15 * Par.Robu -7.56  6.34  1.66  0.10 -2.23 -1.01  0.68 -0.23    F3
 16 * Par.Cras -7.79  4.33  1.42  0.01 -1.80 -0.25  0.04 -0.87    F4
 17 * Megantro -8.23  5.03  1.13 -0.02 -1.41 -0.13 -0.28 -0.13    F5
 18 * Proc.Afr  1.86 -4.28 -2.14 -1.73  2.06  1.80  2.61  2.48    F6
 19 *_
```

Tämän aineiston Andrews-käyrät



syntyvät seuraavalla Survon piirroskaaviolla:

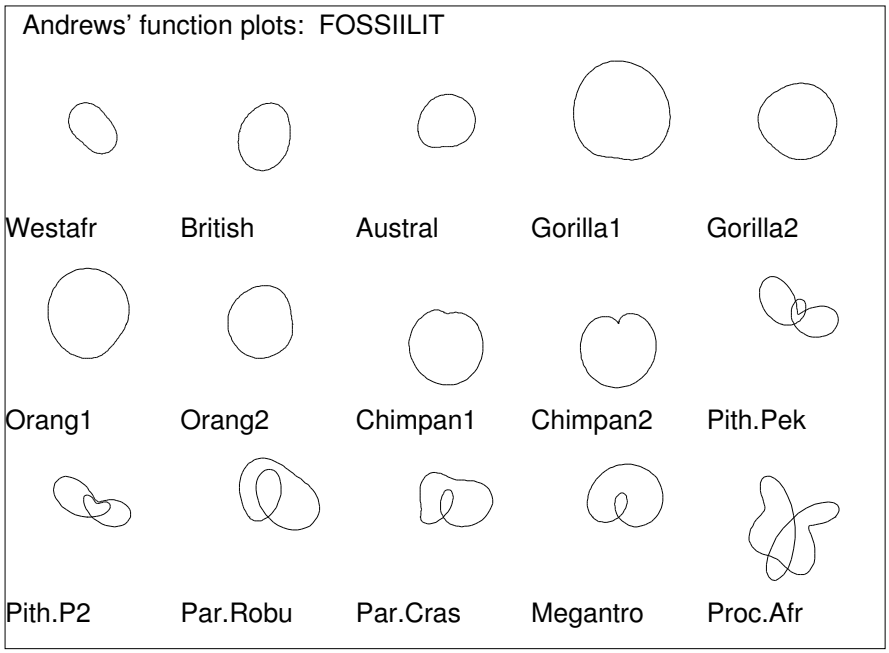
```
 14  1 SURVO 84C EDITOR Sun Jul 24 14:27:16 1994          C:\M\MON\ 150 100 0
 19 *
 20 *GPLOT FOSSIILIT_ / TYPE=ANDREWS LABEL=[Small],Tunnus
 21 *                   YSCALE=-2(1)1
 22 *
 23 *VARIABLES: A        B        Term
 24 *X1         0        1        1/sqrt(2)
 25 *X2         0        1        sin(t)
 26 *X3         0        1        cos(t)
 27 *X4         0        1        sin(2*t)
 28 *X5         0        1        cos(2*t)
 29 *X6         0        1        sin(3*t)
 30 *X7         0        1        cos(3*t)
 31 *X8         0        1        sin(4*t)
 32 *END of plotting specifications
 33 *
```

PLOT-komennossa täsmennys TYPE=ANDREWS tuottaa Andrews-käyrät. Se edellyttää erityistä VARIABLES-luetteloa, joka on tässä riveillä 23-32. Luettelossa kerrotaan muuttujat *X* (tärkeysjärjestyksessä). Jokaista on lupa skaalata muotoon $(X–A)/B$ antamalla parametrit *A* ja *B*. Tässä tapauksessa on A=0 ja B=1 kaikilla muuttujilla eli muuttujanarvoja käytetään sellaisenaan. Eri havaintoja vastaavien käyrien tunnistamiseksi annetaan LABEL-täsmennys. Se ilmoittaa muuttujan, jonka arvoilla jokainen käyristä merkitään sopivasti porrastetuin välein. Havaintotaulukon viimeisenä sarakkeena on muuttuja Tunnus tätä tarkoitusta varten.

Ensimmäisen erottelumuuttujan X1 mukaisesti Andrews-käyrät jakautuvat kahteen ryhmään, joista ylemmässä ovat apinat, alemmassa ihmisrodut ja useimmat fossiilit. Mystinen Proconsul Africanus (F6) kulkee etupäässä apinoitten puolella poiketen kerran ihmisten seuraan.

Kun havaintoja on runsaasti, kuvasta saattaa tulla melko sotkuinen. Jos käyrät piirtää eri kuvina, niitä taas on vaikea erottaa toisistaan. Eräs tapa saada ne helpommin hahmotettaviksi on siirtyä napakoordinaattiesitykseen, jolloin jokaisesta havainnosta muodostuu origon ympärillä kieppuva käyrä. Survossa tällaiset piirrokset syntyvät käyttämällä täsmennystä
TYPE=ANDREWS,POLAR,C .
Tässä C on ylimääräinen additiivinen vakio (minimisäde), joka etäännyttää käyrää origosta ja edelleen parantaa hahmottamista.

Näin piirrettynä fossiiliaineisto näyttää seuraavalta:



Kuva on saatu aikaan kaaviolla:

```
  16  1 SURVO 84C EDITOR Sun Jul 24 18:06:46 1994          C:\M\MON\ 150 100 0
  50 *.....................................................................
  51 *GPLOT FOSSIILIT_ / TYPE=ANDREWS,POLAR,0.2 LABEL=Laji
  52 *                   FSCALING=0,1.5
  53 *
  54 *VARIABLES: A        B        Term
  55 *X1          0        1        1/sqrt(2)
  56 *X2          0        1        sin(t)
  57 *X3          0        1        cos(t)
  58 *X4          0        1        sin(2*t)
  59 *X5          0        1        cos(2*t)
  60 *X6          0        1        sin(3*t)
  61 *X7          0        1        cos(3*t)
  62 *X8          0        1        sin(4*t)
  63 *END of plotting specifications
  64 *
```

Periaatteessa jokainen Andrews-käyrä määrittelee oman aaltomuotonsa, joka on mahdollista realisoida myös äänenä tai sointivärinä. Ei ole tiedossa, onko tällaista koskaan kokeiltu. Huomautettakoon, että Survossa voi kyllä tunnistaa äänen avulla yksittäisen muuttujan havaintosarjasta jaksollisuuksia ja poikkeavia havaintoja. Tämä tapahtuu `FILE SHOW`-operaation yhteydessä.

1.5 Chernoff-naamat

Tässä piirrostavassa, jonka *H.Chernoff* on esittänyt vuonna 1973, muuttujat asetetaan vastaamaan karkeasti piirrettyjen kasvojen eri piirteitä. Survossa on seurattu tarkasti Chernoffin alkuperäistä ehdotusta, jossa valittavia piirteitä oli kaikkiaan 18. Jos aktivoidaan PLOT-komento varustettuna pelkällä TYPE= FACES täsmennyksellä, toimituskenttään kopioituu mallikaavio, jota muokkaamalla soveltaja liittää muuttujat ja kasvojen piirteet toisiinsa. Tämän mallin keskeinen osa on VARIABLES-luettelo:

```
  1  1 SURVO 84C EDITOR Mon Jul 25 09:22:55 1994          C:\M\MON\ 120 100 0
 20 *
 21 *VARIABLES: xmin      xmax     Features                          fmin fmax
 22 *<X1>      <min X1>  <max X1>  Radius_to_corner_of_face_OP        0.6  1.0
 23 *<X2>      <min X2>  <max X2>  Angle_of_OP_to_horizontal         0.0  0.6
 24 *<X3>      <min X3>  <max X3>  Vertical_size_of_face_OU          0.6  1.0
 25 *<X4>      <min X4>  <max X4>  Eccentricity_of_upper_face        0.5  1.5
 26 *<X5>      <min X5>  <max X5>  Eccentricity_of_lower_face        0.5  1.5
 27 *<X6>      <min X6>  <max X6>  Length_of_nose                    0.1  0.5
 28 *<X7>      <min X7>  <max X7>  Vertical_position_of_mouth        0.2  0.8
 29 *<X8>      <min X8>  <max X8>  Curvature_of_mouth_1/R           -4.0  4.0
 30 *<X9>      <min X9>  <max X9>  Width_of_mouth                    0.2  1.0
 31 *<X10>     <min X10> <max X10> Vertical_position_of_eyes         0.0  0.4
 32 *<X11>     <min X11> <max X11> Separation_of_eyes                0.3  0.8
 33 *<X12>     <min X12> <max X12> Slant_of_eyes                    -0.5  0.5
 34 *<X13>     <min X13> <max X13> Eccentricity_of_eyes              0.3  1.0
 35 *<X14>     <min X14> <max X14> Size_of_eyes                      0.1  0.2
 36 *<X15>     <min X15> <max X15> Position_of_pupils               -0.1  0.1
 37 *<X16>     <min X16> <max X16> Vertical_position_of_eyebrows     0.2  0.4
 38 *<X17>     <min X17> <max X17> Slant_of_eyebrows                -0.5  0.5
 39 *<X18>     <min X18> <max X18> Size_of_eyebrows                  0.1  0.5
 40 *END of plotting specifications
 41 *
```

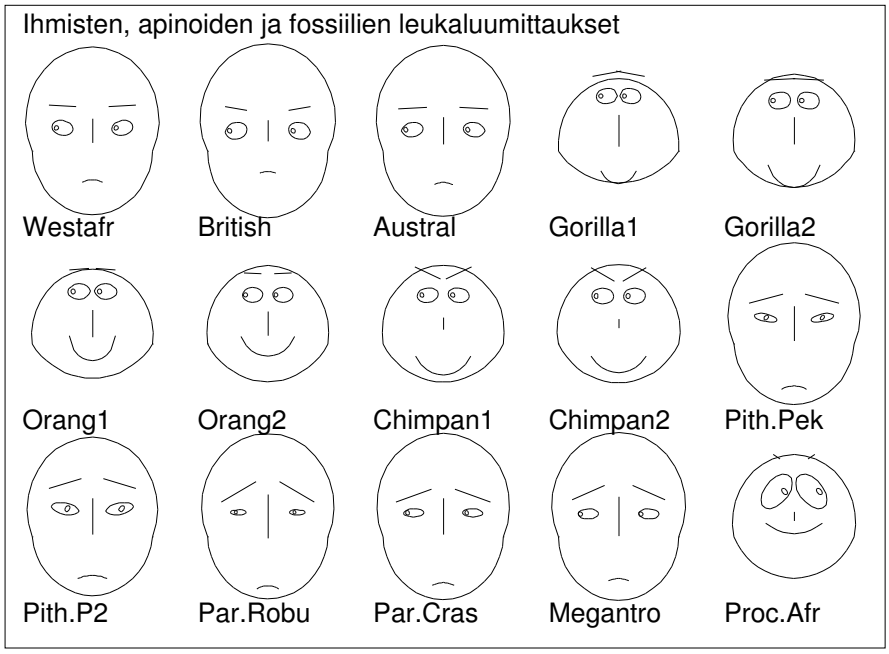
Mallitaulukon viimeisinä sarakkeina ovat kasvojen piirteiden selitykset (Features) ja niiden luonnolliset minimi- ja maksimiarvot. Soveltajan tehtävä on päivittää kolme ensimmäistä saraketta, joissa nimetään eri piirteisiin vaikuttavat muuttujat ja niiden minimi- ja maksimiarvot. Tällöin muuttujien arvot kuvautuvat piirteiksi lineaarisella muunnoksella, joka asettaa minimit minimejä vastaan ja maksimit maksimeja vastaan. Muuttujien minimejä ja maksimeja ei tarvitse erikseen laskea aineistosta, vaan minimin paikalle voi kirjoittaa merkinnän * ja maksimin paikalle ** . Kuvauksen voi kääntää vastakkaiseen suuntaan asettamalla minimin paikalle ** ja maksimin * .

Jokaisen aktivoinnin jälkeen pelkät * ja ** merkinnät korvautuvat todellisilla arvoilla, joiden perässä edelleen on * tai ** . Poistamalla tähdet saadaan kyseinen raja aineistosta riippumattomaksi vakioksi.

Erityisesti kun muuttujia on vähemmän kuin naaman piirteitä, tärkeinä pidettyjä muuttujia kannattaa käyttää useasti. Naaman piirteen voi vakioida (minimin ja maksimin keskiväliin) panemalla muuttujan paikalle merkinnän – .

Fossiiliaineistoa piirrettäessä on houkutus yrittää valita vastaavuudet siten, että ihmisistä ja apinoista tulee jossain määrin itsensä näköisiä. On kuitenkin kohtuutonta kuvitella, että näin saataisiin fossiilit myös näyttämään "oikeilta". Voimme vain havaita, että Chernoffin naamoina useimmat fossiilit ovat

enemmän ihmisen kuin apinan kaltaisia ja että Proconsul Africanus on tässä seurassa tosi outo ilmestys.



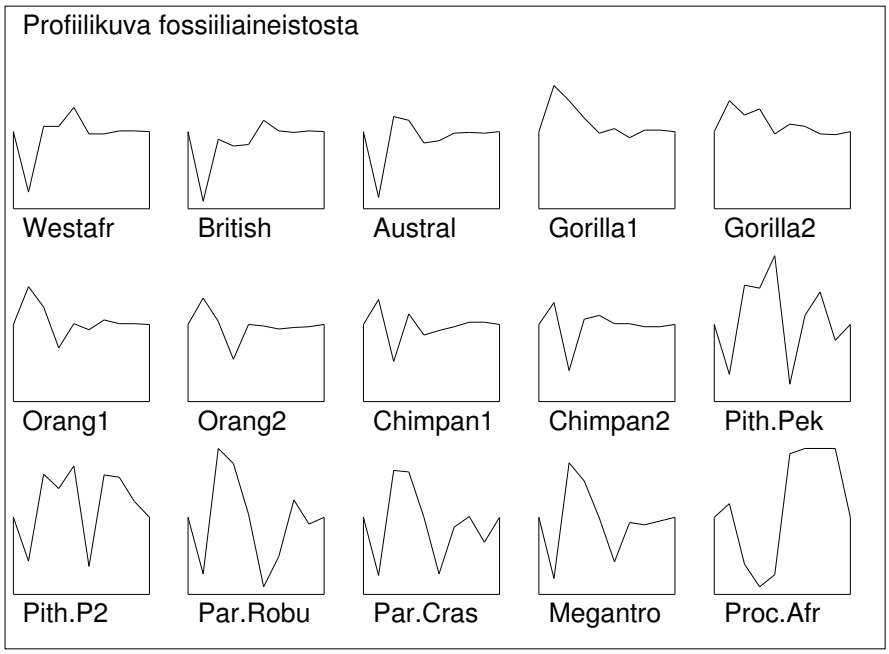
Survon Chernoff-ohjelmassa kuvaruudulla on myös mahdollista käyttää värejä ja esim. "maalata" kasvot ja silmämunat. Tekniikasta on kehitetty myös muita variaatioita. Naamakuvien todellinen hyöty käytännön sovelluksissa on kuitenkin jäänyt vähäiseksi alkuperäisen idean hauskuudesta huolimatta.

1.6 Profiili- ja tähtikuvat

Moniulotteisen aineiston havaintokohtaisia profiileja piirretään Survossa `PLOT`-komennolla, jolla on täsmennys `TYPE=PROFILES`. Tällöin havaintoa $X_1,X_2,...,X_p$ vastaa pisteet $(1,Y_1)$, $(2,Y_2)$, ... (p,Y_p) yhdistävä murtoviiva. Tässä Y:t ovat skaalattuja havaintoarvoja

$$Y_i = X_i/\max(|X_i|) \text{ , } i = 1,2,...,p \text{ .}$$

Fossiiliaineisto näyttää tällöin `GPLOT`-kaaviolla kuvaruutuun piirrettynä seuraavalta:



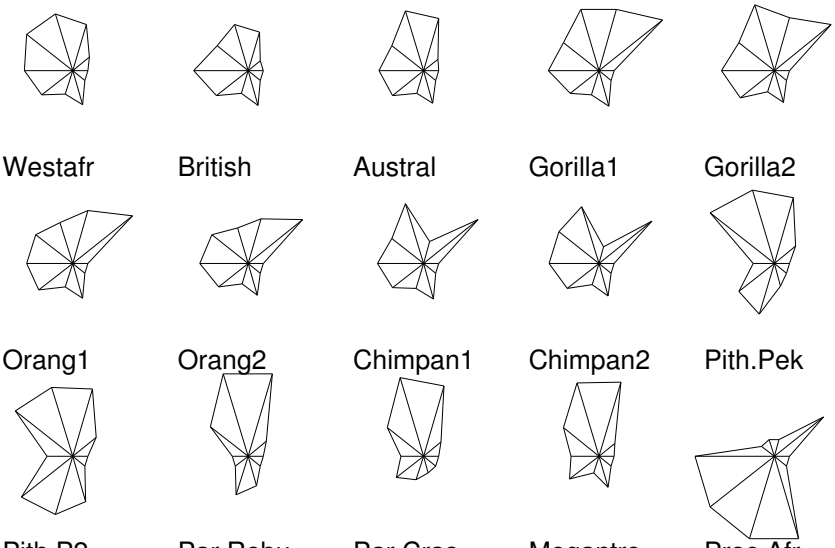
Kuva on saatu aikaan GPLOT-kaaviolla:

```
 16  1 SURVO 84C EDITOR Mon Jul 25 10:39:16 1994          C:\M\MON\ 100 100 0
 19 *
 20 *HEADER=Profiilikuva_fossiiliaineistosta
 21 *GPLOT FOSSIILIT_ /  TYPE=PROFILES  LABEL=Laji
 22 *
```

Tähtikuvissa kutakin havaintoa vastaa origoa kiertävä suljettu murtoviiva siten, että eri muuttujia vastaavat tasavälein suunnatut vektorit. Vierekkäisten vektoreiden kärkipisteet on yhdistetty. Muuttujaa X vastaavan vektorin pituus on $(1-C)(X-\min(X))/[\max(X)-\min(X)]+C$, missä C on vakio (oletusarvona 0.2). Tähtikuva syntyy täsmennyksellä TYPE=STARS,C :

```
 16  1 SURVO 84C EDITOR Mon Jul 25 10:52:37 1994          C:\M\MON\ 100 100 0
 19 *
 20 *HEADER=Tähtikuva_fossiiliaineistosta
 21 *GPLOT FOSSIILIT_ /  TYPE=STARS     LABEL=Laji
 22 *
```

Tähtikuva fossiiliaineistosta



2. Multinormaalijakauma

2.1 Alustavaa johdattelua

Monimuuttujamenetelmissä multinormaalijakaumalla on ehkä vielä keskeisempi asema kuin normaalijakaumalla yhden muuttujan tilastollisissa tarkasteluissa. Multinormaalijakauma on suora normaalijakauman yleistys.

Se voidaan johtaa usealla tavalla. Havainnollisinta on määritellä se toisistaan riippumattomien, normaalijakaumaa noudattavien muuttujien erilaisten painotettujen summien yhteisjakaumana esim. seuraavasti:

Olkoot $Z_1,Z_2,...,Z_p$ riippumattomia, standardoitua normaalijakaumaa $N(0,1)$ noudattavia muuttujia. Muodostetaan uudet muuttujat $X_1,X_2,...,X_p$ Z-muuttujien lineaarisina yhdistelminä

$$
\begin{aligned}
X_1 &= c_{11}Z_1 + c_{12}Z_2 + ... + c_{1p}Z_p + \mu_1 \\
X_2 &= c_{21}Z_1 + c_{22}Z_2 + ... + c_{2p}Z_p + \mu_2 \\
&... \\
X_p &= c_{p1}Z_1 + c_{p2}Z_2 + ... + c_{pp}Z_p + \mu_p
\end{aligned}
$$

eli matriisimuodossa

$$\mathbf{X} = \mathbf{CZ} + \mu$$

missä $\mathbf{X}=(X_1,X_2,...,X_p)$ on X-muuttujien muodostama pystyvektori ja vastaavasti $\mathbf{Z}=(Z_1,Z_2,...,Z_p)$, $\mu=(\mu_1,\mu_2,...,\mu_p)$ sekä $\mathbf{C}$ $p\times p$-kerroinmatriisi.

Muuttujien $X_1,X_2,...,X_p$ yhteisjakaumaa sanotaan multinormaalijakaumaksi ja sen määrittelevät täydellisesti parametrit μ ja $\mathbf{C}$. Itse asiassa tulemme näkemään, että jakauman määrittelemiseksi riittää tuntea odotusarvovektorin μ ohella kovarianssimatriisi $\Sigma = \mathbf{CC}'$.

Multinormaalijakauman syntytapa tulee vielä havainnollisemmaksi käyttämällä hyväksi kerroinmatriisin $\mathbf{C}$ singulaariarvohajotelmaa $\mathbf{C} = \mathbf{UDV}'$, missä $\mathbf{U}$ ja $\mathbf{V}$ ovat $p\times p$-ortogonaalisia matriiseja ja $\mathbf{D}$ (ei-negatiivisten) singulaariarvojen muodostama lävistäjämatriisi.

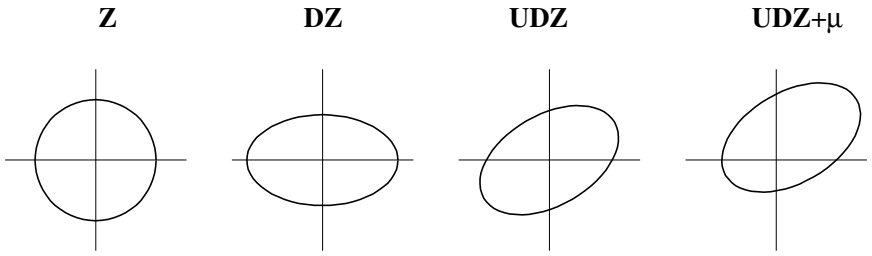
Tällöin

$$\mathbf{X} = \mathbf{CZ} + \mu = \mathbf{UDV}'\mathbf{Z} + \mu \, .$$

Tulemme osoittamaan, että Z-muuttujien ortogonaalinen muunnos (tässä $\mathbf{V}'\mathbf{Z}$) säilyttää muuttujat riippumattomina (0,1)-normaalisina. Näin ollen X-muuttujat voitaisiin määritellä suoraan muodossa

$$\mathbf{X} = \mathbf{UDZ} + \mu \, .$$

Tämä merkitsee, että multinormaalijakauman voi aina ajatella syntyvän (0,1)-normaalisista muuttujista kolmessa vaiheessa. Ensin tehdään muuttujittain venytyksiä ja kutistuksia (**DZ**), sitten kierretään koordinaatistoa (**UDZ**) ja lopuksi siirretään jakauman keskipiste pois origosta (lisäämällä μ).



Tulemme näkemään, että multinormaalijakauman kaikki eriulotteiset reunajakaumatkin ovat (multi)normaalisia. Tämä merkitsee mm. sitä, että multinormaalijakauman em. syntyhistoriassa *Z*-muuttujia voisi olla enemmän kuin lopullisia *X*-muuttujia. Vaikka *Z*-muuttujat eivät olisikaan normaalisia, mutta niitä on "paljon", on osoitettavissa keskeisen raja-arvolauseen tapaan, että *X*-muuttujien yhteisjakauma melko väljin ehdoin lähestyy multinormaalijakaumaa.

Samoin tarkasteltaessa osaa *X*-muuttujista, näiden ehdollinen yhteisjakauma, kun muut *X*-muuttujat asetetaan vakioiksi, on multinormaalinen ja regressiofunktiot (ehdolliset odotusarvot) ovat vakioksi asetettujen *X*-muuttujien lineaarisia funktioita. Tämä viimeinen ominaisuus on myös po. jakauman määritelmän veroinen.

Multinormaalijakauman syntyessä riippumattomien muuttujien lineaaristen yhdistelmien kautta on ilmeistä, että *X*-muuttujien välillä voi vallita vain lineaarisia riippuvuuksia eli korrelaatiokertoimet paljastavat kaiken, mikä koskee muuttujien välisiä riippuvuuksia. Tässä tapauksessa siis korreloimattomuus takaa myös muuttujien riippumattomuuden; seikka, mikä ei välttämättä päde yleisesti moniulotteisissa jakaumissa.

Tämän pohjalta tulee ilmeiseksi, että kaikki multinormaalisuutta edellyttävät tarkastelut saatetaan tehdä muuttujien odotusarvojen, keskihajontojen ja korrelaatiokertoimien avulla. Näiden tunnuslukujen tavanomaiset empiiriset vastineet satunnaisotoksesta laskettuina ovat tyhjentäviä otossuureita eikä esim. korkeamman asteen momentteja tarvita muuta kuin eräissä multinormaalisuutta tutkivissa testeissä.

2.2 Multinormaalijakauman määritelmä ja perusominaisuudet

Tarkennamme äskeistä kuvausta seuraavasti. Olkoot $U_1,U_2,...,U_k$ riippumattomia ja (0,1)-normaalisia satunnaismuuttujia ja $\mathbf{U}=(U_1,U_2,...,U_k)$ niiden muodostama satunnaisvektori. Tällöin odotusarvovektori $E(\mathbf{U})=\mathbf{0}$ ja kovarianssimatriisi $\mathrm{cov}(\mathbf{U})=\mathbf{I}$.

Jokaisen U_i tiheysfunktio on muotoa

$$\phi(u_i) = (2\pi)^{-1/2}\exp(-1/2u_i^2) .$$

Tällöin U-muuttujien riippumattomuuden perusteella satunnaisvektorin $\mathbf{U}$ tiheysfunktio voidaan kirjoittaa näiden komponenttimuuttujien tiheysfunktioiden tulona

$$\begin{aligned} f(\mathbf{u}) = f(u_1,u_2,...,u_k) &= \phi(u_1)\phi(u_2)...\phi(u_k) \\ &= (2\pi)^{-k/2}\exp(-1/2(u_1^2 + u_2^2 + ... + u_k^2)) \\ &= (2\pi)^{-k/2}\exp(-1/2\mathbf{u}'\mathbf{u}) . \end{aligned}$$

Määritellään uusi muuttujavektori $\mathbf{X}=(X_1,X_2,...,X_p)$ lineaarikuvauksella

$$\begin{aligned} X_1 &= c_{11}U_1 + c_{12}U_2 + ... + c_{1k}U_k + \mu_1 \\ X_2 &= c_{21}U_1 + c_{22}U_2 + ... + c_{2k}U_k + \mu_2 \\ &... \\ X_p &= c_{p1}U_1 + c_{p2}U_2 + ... + c_{pk}U_k + \mu_p \end{aligned}$$

eli

(1) $\mathbf{X} = \mathbf{C}\mathbf{U} + \mu$.

Oletetaan, että $p \le k$ ja matriisin $\mathbf{C}$ aste $r(\mathbf{C})=p$. Muussa tapauksessa muuttujat X olisivat lineaarisesti toisistaan riippuvia eikä jakauma olisi aidosti p-ulotteinen.

Muuttujien X odotusarvovektori on

$$E(\mathbf{X}) = \mathbf{C}E(\mathbf{U}) + \mu = \mathbf{C}\cdot\mathbf{0} + \mu = \mu$$

ja kovarianssimatriisi

$$\Sigma = \mathrm{cov}(\mathbf{X}) = E(\mathbf{X}-\mu)(\mathbf{X}-\mu)' = E(\mathbf{C}\mathbf{U}\mathbf{U}'\mathbf{C}') = \mathbf{C}(E\mathbf{U}\mathbf{U}')\mathbf{C}' = \mathbf{C}\mathbf{C}' .$$

Koska $r(\mathbf{C})=p$, on $\Sigma=\mathbf{C}\mathbf{C}'>\mathbf{0}$ (eli positiivisesti definiitti).

Määrätään nyt X-muuttujien yhteisjakauman tiheysfunktio.

Todistetaan ensin apulause:
Olkoot $U_1,U_2,...,U_k$ riippumattomia ja $N(0,1)$. Tällöin myös muuttujat $\mathbf{V}=(V_1,V_2,...,V_k)=\mathbf{Q}\mathbf{U}$, ovat riippumattomia ja $N(0,1)$, jos matriisi $\mathbf{Q}$ on ortogonaalinen (eli $\mathbf{Q}'\mathbf{Q}=\mathbf{Q}\mathbf{Q}'=\mathbf{I}$).

Koska kääntäen $\mathbf{U}=\mathbf{Q}'\mathbf{V}$ ja

$$f_{\mathbf{U}}(\mathbf{u}) = c\exp(-1/2\mathbf{u}'\mathbf{u}),$$

tulee muuttujien V tiheysfunktioksi (sijoittamalla tähän tiheysfunktioon $\mathbf{u}=\mathbf{Q}'\mathbf{v}$ ja kertomalla vastaavalla funktionaalideterminantilla, joka kuvausmatriisin $\mathbf{Q}'$ ortogonaalisuudesta johtuen on 1)

$$f_{\mathbf{V}}(\mathbf{v}) = c\,\exp[-\tfrac{1}{2}(\mathbf{Q}'\mathbf{v})'(\mathbf{Q}'\mathbf{v})]$$
$$= c\,\exp(-\tfrac{1}{2}\mathbf{v}'\mathbf{v}) = f_{\mathbf{U}}(\mathbf{v})\ .$$

Osoitetaan nyt, että jos (1) pätee, on olemassa satunnaisvektori $\mathbf{V}=(V_1,V_2,...,V_p)$, jonka komponentit ovat riippumattomia ja (0,1)-normaalisia siten, että $\mathbf{X}$ voidaan lausua myös niiden avulla muodossa
(2) $\mathbf{X} = \mathbf{A}\mathbf{V} + \mu$,
missä $\mathbf{A}$ on $p\times p$-matriisi ja $\det(\mathbf{A})\neq 0$.

Tämä todistetaan lähtemällä $p\times k$-matriisin $\mathbf{C}$ singulaariarvohajotelmasta $\mathbf{C} = \mathbf{SDT}'$, missä $\mathbf{D}$ on singulaariarvojen $d_1 \geq d_2 \geq ... \geq d_p > 0$ ($r(\mathbf{C})=p$) muodostama lävistäjämatriisi ja $\mathbf{S}$ $p\times p$-ortogonaalinen sekä $\mathbf{T}$ $k\times p$-pystyriveittäin ortogonaalinen eli $\mathbf{T}'\mathbf{T}=\mathbf{I}$. Valitsemalla nyt $\mathbf{V}=\mathbf{T}'\mathbf{U}$ ja $\mathbf{A}=\mathbf{SD}$ saadaan haluttu esitys $\mathbf{X} = \mathbf{C}\mathbf{U} + \mu = \mathbf{SDT}'\mathbf{U} + \mu = \mathbf{A}\mathbf{V} + \mu$. Tässä muuttujat V ovat apulauseen perusteella riippumattomia ja $N(0,1)$, sillä matriisi $\mathbf{T}$ on aina täydennettävissä $k\times k$-ortogonaaliseksi matriisiksi.

Huomattakoon lisäksi, että $\Sigma = \mathrm{cov}(\mathbf{X}) = \mathbf{A}\mathbf{A}' = \mathbf{C}\mathbf{C}'$ ja $\det(\Sigma)=\det(\mathbf{A})^2$.

Koska matriisi $\mathbf{A}$ on säännöllinen, saadaan kääntäen $\mathbf{V} = \mathbf{A}^{-1}(\mathbf{X}-\mu)$ ja

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{V}}(\mathbf{v}(\mathbf{x}))\cdot|\partial(v_1,v_2,...,v_p)/\partial(x_1,x_2,...,x_p)|$$
$$= (2\pi)^{-p/2}\exp(-\tfrac{1}{2}\mathbf{v}'\mathbf{v})\det(\mathbf{A}^{-1})$$
$$= (2\pi)^{-p/2}\det(\Sigma)^{-1/2}\exp[-\tfrac{1}{2}(\mathbf{x}-\mu)'(\mathbf{A}^{-1})'\mathbf{A}^{-1}(\mathbf{x}-\mu)]$$
$$= (2\pi)^{-p/2}\det(\Sigma)^{-1/2}\exp[-\tfrac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)]\ .$$

Siis
$E(\mathbf{X})=\mu$ ja $\mathrm{cov}(\mathbf{X})=\Sigma > \mathbf{0}$
määräävät X-muuttujien yhteisjakauman yksikäsitteisesti.

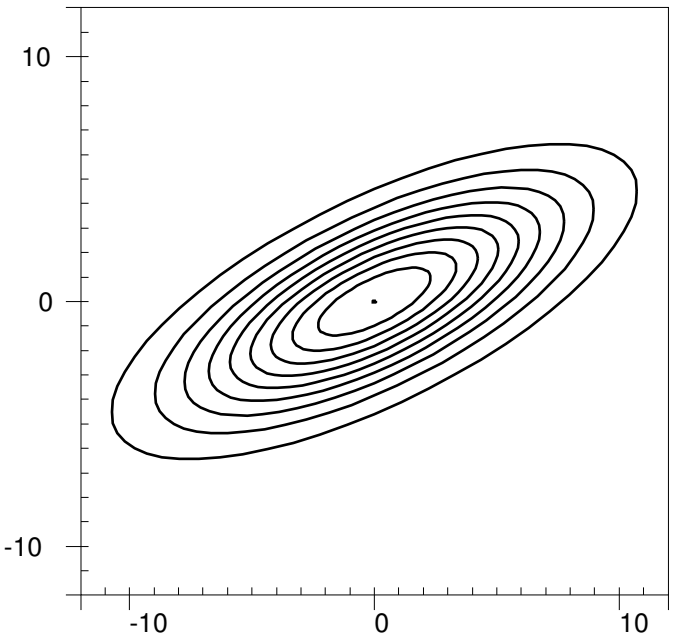
Sanomme, että $\mathbf{X}$ noudattaa p-ulotteista normaalijakaumaa l. multinormaalijakaumaa $N(\mu,\Sigma)$.
Merkitään $\mathbf{X} \sim N(\mu,\Sigma)$, $f_{\mathbf{X}}(\mathbf{x})=n(\mathbf{x} \mid \mu,\Sigma)$, jolloin esim. $\mathbf{V} \sim N(\mathbf{0},\mathbf{I}_p)$.

Kovarianssimatriisin lävistäjällä ovat muuttujien varianssit $\sigma_{11},\sigma_{22},...,\sigma_{pp}$. Näille käytetään myös merkintöjä
$\sigma_{ii}=\sigma_i^2$, $i=1,2,...,p$,
eli $\sigma_1,\sigma_2,...,\sigma_p$ tarkoittavat muuttujien keskihajontoja. Keskihajontojen muodostamaa lävistäjämatriisia merkitään
$\mathbf{D}_\sigma = \mathrm{diag}(\sigma_1,\sigma_2,...,\sigma_p)$,
jolloin muuttujien $\mathbf{X}$ korrelaatiomatriisi $\mathbf{P}$, iso kreikkalainen ρ (rho), saadaan

kaavasta

$$\mathbf{P} = \mathbf{D}_\sigma^{-1}\Sigma\mathbf{D}_\sigma^{-1} .$$

Multinormaalisen satunnaisvektorin **X** tiheysfunktiota hallitsee positiivisesti definiitti neliömuoto $(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)$. Tiheysfunktio on suurimmillaan, kun $\mathbf{x}=\mu$ ja sen arvot vähenevät tästä pisteestä etäännyttäessä siten, että (hyper)-ellipsit eli hajontaellipsit $(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)$ = vakio toimivat tasa-arvokäyrinä.



Kuvassa on sellaisen 2-ulotteisen normaalijakauman tasa-arvokäyriä, jossa muuttujien hajonnat ovat 5 ja 3 sekä korrelaatiokerroin 0.7. Käyrät vastaavat todennäköisyystasoja 0.1,0.2,...,0.9 eli todennäköisyysmassasta 90% on uloimman hajontaellipsin sisällä.

Edelläkäyty tarkastelu osoittaa, että *p*-ulotteinen satunnaisvektori **X** voidaan aina määritellä *p* riippumattoman (0,1)-normaalisen muuttujan avulla.

Annetulla multinormaalisella **X**-vektorilla parametrit μ ja Σ ovat yksikäsitteiset, mutta **V** ja **A** voidaan ajatella valittavaksi useilla tavoilla. Olettaessamme, että satunnaisvektori **X** noudattaa multinormaalijakaumaa $N(\mu,\Sigma)$, emme siis voi tuntea tästä jakaumasta saatujen havaintojen täsmällistä syntytapaa, mutta kaikissa jakauman ominaisuuksia koskevissa tarkasteluissa on lupa käyttää konstruktiota (2), kun vain **A** täyttää ehdon $\Sigma=\mathbf{AA}'$.

Kun siis $\mathbf{X} \sim N(\mu,\Sigma)$, **A**-kuvaus voidaan saada esim. matriisin Σ Cholesky-hajotelmasta $\Sigma=\mathbf{AA}'$, missä **A** on yläkolmiomatriisi tai spektraalihajotelmasta $\Sigma=\mathbf{S}\Lambda\mathbf{S}'$, missä **S** on ortogonaalinen ja Λ ominaisarvojen muodostama lävistäjämatriisi, jolloin $\mathbf{A}=\mathbf{S}\Lambda^{1/2}$.

Edellä on oletettu matriisi **A** täysiasteiseksi, jolloin sillä ja kovarianssimatriisilla Σ on käänteismatriisi. Tällöin jakauma on aidosti p-ulotteinen ja sille voidaan kirjoittaa edellä todettu tiheysfunktion lauseke.

Voimme jo johdannossa mainitulla tavalla vielä yksinkertaistaa määritelmää (2) matriisin **A** singulaariarvohajotelman **A=SDT**' avulla. Tällöin

$$\mathbf{X} = \mathbf{AV} + \mu = \mathbf{SDT}'\mathbf{V} + \mu = \mathbf{SDW} + \mu$$

eli

(3) $\mathbf{X} = \mathbf{SDW} + \mu$,

missä $\mathbf{W} \sim N(\mathbf{0},\mathbf{I})$ edellä olevan apulauseen nojalla, **D** on positiivisten singulaariarvojen $d_1 \geq d_2 \geq ... \geq d_p > 0$ muodostama lävistäjämatriisi ja **S** $p \times p$-ortogonaalinen matriisi.

Kuten myöhemmin tulemme näkemään, muuttujat $\mathbf{DW}=(d_1W_1,...,d_pW_p)$ ovat muuttujien **X** pääkomponentteja, joiden voimakkuuksia (itse asiassa keskihajontoja ja geometrisesti hajontaellipsoidien pääakseleiden pituuksia) vastaavat singulaariarvot.

Esittämämme konstruktiivinen määritelmä antaisi mahdollisuuden käsitellä vaivatta myös vajaa-asteisia tapauksia, joissa osa singulaariarvoista on nollia, mutta jatkossa tarkastelemme lähes poikkeuksetta vain täysiulotteista multinormaalijakaumaa.

Tutkiessamme multinormaalijakauman ominaisuuksia käytämme usein apuna konstruktiivisia määritelmiä (1), (2) ja (3), jotka yleensä tekevät tarkastelut yksinkertaisemmiksi kuin jos perustaisimme ne multinormaalijakauman tiheysfunktion esitykseen. Useimmat oppikirjat lähtevät liikkeelle suoraan esim. tiheysfunktiosta tai karakteristisesta funktiosta, jolloin helposti kadotetaan jakauman luonnollinen tausta.

2.2.1 Reunajakaumat

Tulemme useasti tarkastelemaan p komponentin satunnaisvektoria **X** kahden osavektorin $\mathbf{X}^{(1)}$ ja $\mathbf{X}^{(2)}$ yhdistelmänä siten, että $\mathbf{X}^{(1)}$ käsittää q ($q<p$) ensimmäistä muuttujaa $\mathbf{X}^{(1)}=(X_1,X_2,...,X_q)$ ja $\mathbf{X}^{(2)}$ loput p-q muuttujaa $\mathbf{X}^{(2)}=(X_{q+1},X_{q+2},...,X_p)$. Mikä tahansa muuttujien osajoukko saadaan näiden tarkastelujen piiriin järjestämällä muuttujavektorin **X** komponentit sopivasti uudelleen.

Ositettujen matriisien merkintätapoja noudattaen on siis

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \\ \mathbf{X}^{(2)} \end{bmatrix},$$

jolloin odotusarvovektorin μ ja kovarianssimatriisin Σ ositetut esitykset ovat

$$\mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11}\ \Sigma_{12} \\ \Sigma_{21}\ \Sigma_{22} \end{bmatrix}.$$

Osoitamme nyt, että muuttujavektorin $\mathbf{X}^{(1)}$ jakauma on $N(\mu^{(1)},\Sigma_{11})$. Tämä tapahtuu määritelmän (2) avulla eli kirjoittamalla $\mathbf{X} = \mathbf{AV} + \mu$ ositetussa muodossa

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix} \mathbf{V} + \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix},$$

jolloin

$\mathbf{X}^{(1)} = \mathbf{A}_1\mathbf{V} + \mu^{(1)}$.

Tällöin määritelmän (1) mukaan

$\mathbf{X}^{(1)} \sim N(\mu^{(1)}, \mathbf{A}_1\mathbf{A}_1') = N(\mu^{(1)},\Sigma_{11})$.

2.2.2 Muuttujien vaihto

Konstruktiivisen määritelmän mukaan on mitä ilmeisintä, että multinormaalisuus säilyy muuttujien lineaarisissa kuvauksissa. Näytämme täsmällisemmin, että jos $\mathbf{X} \sim N(\mu,\Sigma)$ ja $\mathbf{Y}=\mathbf{BX}$, missä $\mathbf{B}$ on täysiasteinen $m\times p$-matriisi ($r(\mathbf{B})=m$, $m\le p$), niin $\mathbf{Y} \sim N(\mathbf{B}\mu,\mathbf{B}\Sigma\mathbf{B}')$.

Tämän todistamiseksi käytämme määritelmää (2) eli $\mathbf{X}=\mathbf{AV}+\mu$, jolloin

$\mathbf{Y} = \mathbf{BX} = \mathbf{BAV} + \mathbf{B}\mu$

eli $\mathbf{Y}$ syntyy määritelmän (1) mukaan (0,1)-normaalisista V-muuttujista käyttäen kuvausmatriisia $\mathbf{BA}$ ja lisäystä $\mathbf{B}\mu$. Siis $\mathbf{Y} \sim N(\mathbf{B}\mu,\mathbf{BAA}'\mathbf{B}')$ eli $\mathbf{Y} \sim N(\mathbf{B}\mu,\mathbf{B}\Sigma\mathbf{B}')$, sillä $\mathbf{AA}'=\Sigma$.

Erityisesti havaitaan, että jokainen $\mathbf{X}$-muuttujien lineaarinen kombinaatio noudattaa tavallista yksiulotteista normaalijakaumaa seuraavasti. Olkoon $\alpha=(\alpha_1,\alpha_2,...,\alpha_p)$ p komponentin pystyvektori. Tällöin

$$Y = \alpha_1X_1 + \alpha_2X_2 + ... + \alpha_pX_p = \alpha'\mathbf{X} \sim N(\alpha'\mu,\ \alpha'\Sigma\alpha) .$$

2.2.3 Ehdolliset jakaumat

Ehdolliset jakaumat johdetaan tavallisesti operoimalla yhteisjakauman ja reunajakaumien tiheysfunktioiden avulla. Konstruktiivinen määritelmämme tarjoaa tässäkin periaatteessa yksinkertaisemman ja selvemmän tavan.

Tarkastellaan osavektorin $\mathbf{X}^{(1)}$ jakaumaa ehdolla $\mathbf{X}^{(2)}=\mathbf{c}^{(2)}$. Toimitaan määritelmän (2) $\mathbf{X}=\mathbf{AV}+\mu$ pohjalta ja valitaan $\mathbf{A}$ kovarianssimatriisin Σ Cholesky-hajotelmasta $\Sigma=\mathbf{AA}'$, jossa $\mathbf{A}$ on yläkolmiomatriisi.

Osittamalla vektori $\mathbf{V}$ samalla tavalla kuin $\mathbf{X}$ saadaan

$$\mathbf{X}=\begin{bmatrix}\mathbf{X}^{(1)}\\ \mathbf{X}^{(2)}\end{bmatrix}=\begin{bmatrix}\mathbf{A}_{11}\mathbf{A}_{12}\\ \mathbf{0}\quad\mathbf{A}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{V}^{(1)}\\ \mathbf{V}^{(2)}\end{bmatrix}+\begin{bmatrix}\mu^{(1)}\\ \mu^{(2)}\end{bmatrix},$$

jolloin

$$\mathbf{X}^{(2)}=\mathbf{A}_{22}\mathbf{V}^{(2)}+\mu^{(2)}=\mathbf{c}^{(2)} \text{ eli } \mathbf{V}^{(2)}=\mathbf{A}_{22}^{-1}(\mathbf{c}^{(2)}-\mu^{(2)})$$

ja

$$\begin{aligned}\mathbf{X}^{(1)}&=\mathbf{A}_{11}\mathbf{V}^{(1)}+\mathbf{A}_{12}\mathbf{V}^{(2)}+\mu^{(1)}\\ &=\mathbf{A}_{11}\mathbf{V}^{(1)}+\mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{c}^{(2)}-\mu^{(2)})+\mu^{(1)}.\end{aligned}$$

Määritelmän (1) perusteella toteamme, että $\mathbf{X}^{(1)}$ ehdolla $\mathbf{X}^{(2)}=\mathbf{c}^{(2)}$ noudattaa multinormaalijakaumaa

$$N(\mu^{(1)}+\mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{c}^{(2)}-\mu^{(2)}),\ \mathbf{A}_{11}\mathbf{A}_{11}').$$

Kirjoittamalla yhtälö $\Sigma=\mathbf{AA}'$ ositetussa muodossa

$$\Sigma=\begin{bmatrix}\Sigma_{11}\ \Sigma_{12}\\ \Sigma_{21}\ \Sigma_{22}\end{bmatrix}=\begin{bmatrix}\mathbf{A}_{11}\ \mathbf{A}_{12}\\ \mathbf{0}\quad\mathbf{A}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{A}'_{11}\ \mathbf{0}\\ \mathbf{A}'_{12}\ \mathbf{A}'_{22}\end{bmatrix}$$

näytetään "helposti", että

$$\mathbf{A}_{11}\mathbf{A}_{11}'=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

ja

$$\mathbf{A}_{12}\mathbf{A}_{22}^{-1}=\Sigma_{12}\Sigma_{22}^{-1}.$$

Näin ollen $\mathbf{X}^{(1)}$ ehdolla $\mathbf{X}^{(2)}=\mathbf{c}^{(2)}$ noudattaa multinormaalijakaumaa

$$N(\mu^{(1)}+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{c}^{(2)}-\mu^{(2)}),\ \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

Matriisista

$$\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$

jota sanotaan jäännös- tai osittaiskovarianssimatriisiksi, käytetään tavallisesti lyhennysmerkintää

$$\Sigma_{11.2}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

On huomionarvoista, että $\Sigma_{11.2}$ ei riipu lainkaan siitä, mille tasolle $\mathbf{c}^{(2)}$ ehtomuuttujat $\mathbf{X}^{(2)}$ on asetettu. Tällöin on oikeutettua sanoa, että $\Sigma_{11.2}$ kertoo kovarianssimatriisina muuttujien $\mathbf{X}^{(1)}$ välisistä riippuvuuksista, kun muuttujien $\mathbf{X}^{(2)}$ vaikutus on poistettu.

Toinen merkittävä seikka on ehdollisten odotusarvojen eli regressiofunktioiden

$$\mathrm{E}(\mathbf{X}^{(1)} \mid \mathbf{X}^{(2)}=\mathbf{c}^{(2)}) = \mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{c}^{(2)} - \mu^{(2)})$$

rakenne. Ne riippuvat ehtomuuttujien arvoista $\mathbf{c}^{(2)}$ vain lineaarisesti. Tämä on ominaisuus, joka on voimassa vain multinormaalijakaumassa.

$q\times q$-osittaiskovarianssimatriisin $\Sigma_{11.2}$ alkioita merkitään

$$\sigma_{ij.\ q+1,\ldots,p},\ i,j=1,2,\ldots,q\ ,$$

jolloin ehdollisen jakauman korrelaatiokertoimet eli osittaiskorrelaatiokertoimet ovat

$$\rho_{ij.\ q+1,\ldots,p} = \sigma_{ij.\ q+1,\ldots,p} \,/\, \sqrt{\sigma_{ii.\ q+1,\ldots,p}\sigma_{jj.\ q+1,\ldots,p}}\ ,\quad i,j=1,2,\ldots,q\ .$$

Esim. 1

Tarkastellaan tapausta $p=2$, $q=1$ eli kaksiulotteista normaalijakaumaa ja ensimmäisen muuttujan X_1 jakaumaa ehdolla, että toinen muuttuja X_2 on vakio x_2. Tällöin

$$\mu^{(1)}=\mu_1,\ \mu^{(2)}=\mu_2,\ \Sigma_{11}=\sigma_1^2,\ \Sigma_{12}=\sigma_1\sigma_2\rho,\ \Sigma_{22}=\sigma_2^2,$$

missä ρ on muuttujien välinen korrelaatiokerroin. Edelleen

$$\Sigma_{11.2} = \sigma_1^2 - \sigma_1\sigma_2\rho\sigma_2^{-2}\sigma_1\sigma_2\rho = \sigma_1^2(1-\rho^2)$$

eli kyseinen jakauma on

$$N(\mu_1+\sigma_1\rho/\sigma_2(x_2-\mu_2),\ \sigma_1^2(1-\rho^2))\ .$$

Esim. 2: "Kopiointiprosessi"

Mitataan tietyn kappaleen pituus X_1 ja yritetään tehdä siitä kopio, mutta kopioinnissa syntyy satunnaisvirhettä niin, että kopion pituudeksi tulee X_2. Jatketaan kopioimalla kopio, jolloin saadaan kappale, jonka pituus on X_3, ja toistetaan kopioiden kopioimista niin, että saadaan kaiken kaikkiaan mittaustulokset $\mathbf{X}=(X_1,X_2,\ldots,X_p)$.

Oletetaan, että $\mathbf{X}$ noudattaa multinormaalijakaumaa ja että peräkkäisten "sukupolvien" X_i ja X_{i+1} välinen korrelaatiokerroin on vakio ρ kaikilla arvoilla $i=1,2,\ldots,p$-1. Kopioinnin luonteen perusteella oletetaan lisäksi, että kaikki tieto pituudesta "sukupolvien" i ja i+2 välillä kulkee sukupolven i+1 kautta. Siis vaaditaan, että $\rho_{i,i+2.\ i+1}=0$ kaikilla $i=1,2,\ldots,p$-2.

Katsotaan, mitä tästä voi päätellä yleensä korrelaatiomatriisin rakenteen suhteen. Koska korrelaatiokertoimet ja osittaiskorrelaatiokertoimet säilyvät muuttujien erillisissä, positiiviskertoimisisissa lineaarisissa muunnoksissa, riittää yksinkertaisuuden vuoksi tarkastella tilannetta, jossa muuttujien keskihajonnat ovat 1 eli $\Sigma=\mathbf{P}$.

Rajoitutaan tapaukseen p=3, jolloin korrelaatiomatriisi P on muotoa

$$\mathrm{P}=\begin{array}{c}\begin{matrix}1 & 2 & 3\end{matrix}\\ \begin{bmatrix}1 & \rho & x\\ \rho & 1 & \rho\\ x & \rho & 1\end{bmatrix}\end{array}\begin{matrix}1\\2\\3\end{matrix} \qquad \mathrm{P}^{*}=\begin{array}{c}\begin{matrix}1 & 3 & 2\end{matrix}\\ \begin{bmatrix}1 & x & \rho\\ x & 1 & \rho\\ \rho & \rho & 1\end{bmatrix}\end{array}\begin{matrix}1\\3\\2\end{matrix}$$

Tässä x:llä on merkitty tuntematonta korrelaatiokerrointa $\rho_{13}=\rho_{31}$ ja osittaiskovarianssin $\sigma_{13.\,2}$ laskemisen helpottamiseksi on vaihdettu muuttujien X_2 ja X_3 paikat, jolloin saadaan yllä toisena oleva korrelaatiomatriisi P^{*}. Kun siis muuttujan X_2 vaikutus poistetaan, saadaan osittaiskovarianssimatriisiksi muuttujien X_1 ja X_3 välillä

$$\begin{bmatrix}1 & x\\ x & 1\end{bmatrix}-\begin{bmatrix}\rho\\ \rho\end{bmatrix}\begin{bmatrix}1\end{bmatrix}^{-1}\begin{bmatrix}\rho & \rho\end{bmatrix}=\begin{bmatrix}1-\rho^2 & x-\rho^2\\ x-\rho^2 & 1-\rho^2\end{bmatrix},$$

jolloin vaatimuksesta $\sigma_{13.2}=x-\rho^2=0$ seuraa $x=\rho^2$. Harjoitustehtäväksi jää osoittaa, että yleisesti $\rho_{ij}=\rho^{|\,i-j\,|}$, $i,j=1,2,...,p$.

Harjoitustehtäväksi jätetään myös selvittää, miten matriisi **A** tulisi valita, jotta määritelmä (2) $\mathbf{X}=\mathbf{AV}+\mu$ antaa muuttujavektorille **X** juuri edellä esitetyn korrelaatiorakenteen.

2.2.4 Muuttujaryhmien riippumattomuus

Muuttujaryhmät $\mathbf{X}^{(1)}$ ja $\mathbf{X}^{(2)}$ ovat toisistaan riippumattomia vain jos niiden syntyyn vaikuttavat eri V-muuttujat konstruktiivisessa määritelmässä (2) $\mathbf{X}=\mathbf{AV}+\mu$. On siis voimassa esim.

$$\mathbf{X}^{(1)}=\mathbf{A}_{11}\mathbf{V}^{(1)}+\mu^{(1)},$$
$$\mathbf{X}^{(2)}=\mathbf{A}_{22}\mathbf{V}^{(2)}+\mu^{(2)}$$

eli $\mathbf{A}_{12}=\mathbf{0}$ ja $\mathbf{A}_{21}=\mathbf{0}$. Tällöin

$$\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}=\begin{bmatrix}\mathbf{A}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{A}'_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}'_{22}\end{bmatrix}=\begin{bmatrix}\mathbf{A}_{11}\mathbf{A}'_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_{22}\mathbf{A}'_{22}\end{bmatrix}$$

eli $\Sigma_{12}=\mathbf{0}$ (samoin kuin $\Sigma_{21}=\mathbf{0}$) ja muuttujaryhmien väliset korrelaatiokertoimet ovat nollia.

Kääntäen, jos $\Sigma_{12}=\mathbf{0}$, $\mathbf{X}^{(1)}$ ehdolla $\mathbf{X}^{(2)}=\mathbf{c}^{(2)}$ on $N(\mu^{(1)},\Sigma_{11})$, mikä on $\mathbf{X}^{(1)}$:n reunajakauma. Muuttujaryhmät ovat tällöin siis myös riippumatttomia. On huomattava, että korreloimattomuudesta ei yleisesti seuraa riippumattomuutta. Tämä ominaisuus koskee vain multinormaalijakaumaa, jossa korreloi-

mattomuus (eli lineaarinen riippumattomuus) ja yleinen riippumattomus ovat ekvivalentteja.

2.2.5 Muuttujaryhmien riippuvuus

Jos $\Sigma_{12}\neq\mathbf{0}$, muuttujaryhmien $\mathbf{X}^{(1)}$ ja $\mathbf{X}^{(2)}$ välillä on riippuvuuksia, jotka edellä todetun perusteella voivat olla luonteeltaan vain lineaarisia ja ilmaistavissa korrelaatiokertoimien avulla. Tätä riippuvuutta kuvaavat tehokkaimmin kanoniset korrelaatiot ja tapauksessa q=1 yhteiskorrelaatiokerroin.

Tavoitteenamme on tässä tarkastella yhteiskorrelaatiokerrointa, mutta aloitamme yleisesti tapauksesta, jossa q ei ole välttämättä 1 eli kummassakin muuttujaryhmässä on useita komponenttimuuttujia. Yritämme etsiä mahdollisimman hyvää riippuvuutta muuttujaryhmien $\mathbf{X}^{(1)}$ ja $\mathbf{X}^{(2)}$ välillä määrittelemällä kummassakin yleisen lineaarisen yhdistelmän

$$\alpha'\mathbf{X}^{(1)} = \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_q X_q ,$$
$$\beta'\mathbf{X}^{(2)} = \beta_1 X_{q+1} + \beta_2 X_{q+2} + ... + \beta_p X_p$$

siten, että näiden yhdistettyjen muuttujien välinen korrelaatiokerroin tulee mahdollisimman suureksi. Voimme yleisyyttä loukkaamatta olettaa, että odotusarvot ovat nollia ja että α- ja β-kertoimet on normeerattu siten, että yhdistettyjen muuttujien varianssit ovat ykkösiä.

Tällöin korrelaatiokertoimen maksimointi tarkoittaa lausekkeen

$$\mathrm{cov}(\alpha'\mathbf{X}^{(1)},\beta'\mathbf{X}^{(2)}) = \mathrm{E}(\alpha'\mathbf{X}^{(1)}\mathbf{X}^{(2)'}\beta) = \alpha'\Sigma_{12}\beta$$

maksimointia kerroinvektoreiden α ja β suhteen ehdoilla

$$\mathrm{var}(\alpha'\mathbf{X}^{(1)}) = \mathrm{E}(\alpha'\mathbf{X}^{(1)}\mathbf{X}^{(1)'}\alpha) = \alpha'\Sigma_{11}\alpha = 1 ,$$
$$\mathrm{var}(\beta'\mathbf{X}^{(2)}) = \mathrm{E}(\beta'\mathbf{X}^{(2)}\mathbf{X}^{(2)'}\beta) = \beta'\Sigma_{22}\beta = 1 .$$

Otamme käyttöön Cholesky-hajotelmat

$$\Sigma_{11}=\mathbf{S}_1'\mathbf{S}_1 \text{ ja } \Sigma_{22}=\mathbf{S}_2'\mathbf{S}_2$$

sekä uudet vektorit $\mathbf{u}=\mathbf{S}_1\alpha$ ja $\mathbf{v}=\mathbf{S}_2\beta$. Tällöin tehtävämme muuntuu lausekkeen

$$\mathbf{u}'(\mathbf{S}_1^{-1})'\Sigma_{12}\mathbf{S}_2^{-1}\mathbf{v} = \mathbf{u}'\mathbf{A}\mathbf{v}$$

maksimoinniksi ehdoilla $\mathbf{u}'\mathbf{u}=\mathbf{v}'\mathbf{v}$=1. Tässä on merkitty lyhyyden vuoksi

$$\mathbf{A} = (\mathbf{S}_1^{-1})'\Sigma_{12}\mathbf{S}_2^{-1}$$

ja $\mathbf{A}$ on muodoltaan $q\times(p-q)$-matriisi.

Kuten liitteessä 2 osoitetaan, yleisesti $\mathbf{u}'\mathbf{A}\mathbf{v}$ maksimoituu ehdoilla $\mathbf{u}'\mathbf{u}=\mathbf{v}'\mathbf{v}$=1, kun matriisin $\mathbf{A}$ singulaariarvohajotelmasta $\mathbf{A}=\mathbf{U}\mathbf{D}\mathbf{V}'$ valitaan suurin singulaariarvo d_1 sekä tätä vastaavat (ensimmäiset) pystyvektorit $\mathbf{u}^{(1)}$ ja $\mathbf{v}^{(1)}$ ortogonaalisista matriiseista $\mathbf{U}$ ja $\mathbf{V}$. Singulaariarvo d_1 on samalla lausekkeen $\mathbf{u}'\mathbf{A}\mathbf{v}$ maksimiarvo eli suurin mahdollinen korrelaatiokerroin. Sitä sanotaan

ensimmäiseksi kanoniseksi korrelaatiokertoimeksi. Tähän palataan yleisesti myöhemmin kanonisen analyysin yhteydessä.

Nyt kiinnostaa lähinnä tapaus q=1, jolloin puhutaan muuttujan X_1 ja muuttujaryhmän $\mathbf{X}^{(2)}$ yhteiskorrelaatiokertoimesta. Se on analoginen lineaarisen regressioanalyysin yhteiskorrelaatiokertoimen kanssa. Tällöin matriisi $\mathbf{A}$ on p-1 komponentin vaakavektori

$$\mathbf{A} = \sigma_1^{-1}\Sigma_{12}\mathbf{S}_2^{-1}$$

ja sen singulaariarvohajotelma surkastuu muotoon

$$\mathbf{A} = 1 \cdot d_1 \cdot \mathbf{v}^{(1)'}$$

Koska $\mathbf{v}^{(1)'}\mathbf{v}^{(1)}$=1, niin d_1^2 on yksinkertaisesti $\mathbf{AA}'$ eli

$$d_1^2 = \mathbf{AA}' = \sigma_1^{-1}\Sigma_{12}\mathbf{S}_2^{-1}(\mathbf{S}_2^{-1})'\Sigma_{21}\sigma_1^{-1} = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} / \sigma_1^2.$$

Tällöin muuttujan X_1 yhteiskorrelaatiokerroin muuttujien $X_2,...,X_p$ suhteen on

$$R_{1.\,23...p} = d_1 = \sqrt{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} \;/\; \sigma_1 .$$

Herää kysymys, mikä on se kerroinvektori β, joka antaa tämän maksimikorrelaation. Todetaan suoraan edellisten tulosten perusteella, että

$$\beta = \mathbf{S}_2^{-1}\mathbf{v}^{(1)} \propto \mathbf{S}_2^{-1}(\mathbf{S}_2^{-1})'\Sigma_{12}' = \Sigma_{22}^{-1}\Sigma_{12}'$$

eli

$$\beta'\mathbf{x}^{(2)} = \Sigma_{12}\Sigma_{22}^{-1}\mathbf{x}^{(2)}$$

on X_1:n ehdollinen odotusarvo l. regressiofunktio, kun $\mathbf{X}^{(2)}$=$\mathbf{x}^{(2)}$ ja otetaan huomioon, että odotusarvot on oletettu nolliksi.

Yhteiskorrelaatiokerroin $R_{1.\,23...p}$ voidaan lausua usein eri tavoin, jotka ilmenevät seuraavasta harjoitustehtävästä:

Osoita, että ositetun Σ-matriisin determinantti voidaan esittää muodossa

$$|\Sigma| = |\Sigma_{22}|\,|\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}| = |\Sigma_{22}|\,|\Sigma_{11.2}|$$

ja tämän perusteella, että

$$1 - R_{1.\,23...p}^2 = \frac{|\Sigma|}{\sigma_{11}|\Sigma_{22}|},$$

$$\sigma_{11.\,23...p} = \sigma_{11}(1 - R_{1.\,23...p}^2) \text{ ja}$$

$$R_{1.\,23...p} = \sqrt{1 - 1/\rho^{11}} \;,$$

kun korrelaatiomatriisin käänteismatriisin alkioita merkitään $\mathbf{P}^{-1}$=$[\rho^{ij}]$.

2.2.6 Karakteristinen funktio

Satunnaisvektorin $\mathbf{X}=(X_1,X_2,...,X_p)$ karakteristinen funktio (cf) määritellään muodossa

$$\phi_{\mathbf{X}}(\mathbf{t}) = \phi(\mathbf{t}) = \mathrm{E}(e^{i\mathbf{t}'\mathbf{X}}) = \mathrm{E}(\cos\mathbf{t}'\mathbf{X}) + i\mathrm{E}(\sin\mathbf{t}'\mathbf{X}) ,$$

missä $\mathbf{t}=(t_1,t_2,...,t_p)$ on reaalinen vektori. Jos muuttujat $X_1,X_2,...,X_p$ ovat riippumattomia,

$$\phi(\mathbf{t}) = \mathrm{E}(\exp(i(t_1X_1+...+t_pX_p))) = \mathrm{E}(\exp(it_1X_1))\cdots\mathrm{E}(\exp(it_pX_p)) = \phi_{X_1}(t_1)\cdots\phi_{X_p}(t_p) .$$

Johdamme nyt multinormaalijakauman $N(\mu,\Sigma)$ karakteristisen funktion sen tiedon pohjalta, että $N(0,1)$-muuttujan karakteristinen funktio on

$$\phi(u) = e^{-\frac{1}{2}u^2} .$$

Multinormaalijakauman konstruktiivisen määritelmän (2) mukaan satunnaisvektori $\mathbf{X}$ voidaan lausua muodossa

$$\mathbf{X} = \mathbf{AV} + \mu,$$

missä muuttujat $\mathbf{V}=(V_1,V_2,...,V_p)$ ovat $N(0,1)$-jakautuneita ja riippumattomia. sekä $\mathbf{AA}'=\Sigma$. Muuttujavektorin $\mathbf{V}$ karakteristinen funktio on

$$\phi_{\mathbf{V}}(s) = \phi_{V_1}(s_1)\cdots\phi_{\mathbf{V}_p}(s_p) = e^{-\frac{1}{2}s_1^2}\cdots e^{-\frac{1}{2}s_p^2} = e^{-\frac{1}{2}\mathbf{s}'\mathbf{s}} .$$

Tällöin

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathrm{E}(e^{i\mathbf{t}'\mathbf{X}}) = \mathrm{E}(e^{i\mathbf{t}'(\mathbf{AV}+\mu)}) = e^{i\mathbf{t}'\mu}\mathrm{E}(e^{i(\mathbf{A}'\mathbf{t})'\mathbf{V}}) = e^{i\mathbf{t}'\mu}e^{-\frac{1}{2}(\mathbf{A}'\mathbf{t})'(\mathbf{A}'\mathbf{t})}$$

eli

$$\phi_{\mathbf{X}}(\mathbf{t}) = e^{i\mathbf{t}'\mu-\frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}} .$$

Karakteristisen funktion avulla on mahdollista todistaa monia multinormaalijakauman keskeisiä ominaisuuksia. Näistä eräs merkittävimpiä on se, että vain multinormaalijakaumassa kaikki muuttujien lineaarikombinaatiot ovat normaalisia:

Oletetaan, että $\mathbf{Y}$ on satunnaisvektori, jolla $\mathbf{s}'\mathbf{Y}$ on normaalinen jokaisella vektorilla $\mathbf{s}$. Olkoon $\mathrm{E}(\mathbf{Y})=\mu$ ja $\mathrm{cov}(\mathbf{Y})=\Sigma$. Tällöin $\mathrm{E}(\mathbf{s}'\mathbf{Y})=\mathbf{s}'\mu$ ja $\mathrm{var}(\mathbf{s}'\mathbf{Y})=\mathbf{s}'\Sigma\mathbf{s}$. Satunnaismuuttujan $\mathbf{s}'\mathbf{Y}$ karakteristinen funktio on

$$\phi_{\mathbf{s}'\mathbf{Y}}(t) = \mathrm{E}(e^{it\mathbf{s}'\mathbf{Y}}) = e^{it\mathbf{s}'\mu-\frac{1}{2}t^2\mathbf{s}'\Sigma\mathbf{s}} .$$

Kun yksinkertaisesti asetetaan $t=1$, saadaan

$$\mathrm{E}(e^{i\mathbf{s}'\mathbf{Y}}) = e^{i\mathbf{s}'\mu-\frac{1}{2}\mathbf{s}'\Sigma\mathbf{s}} = \phi_{\mathbf{Y}}(\mathbf{s})$$

eli

$$\mathbf{Y} \sim N(\mu,\Sigma) .$$

Tämän tuloksen käytännölliseen merkitykseen palaamme aivan kohta. Tässä kuitenkin toteamme ensin, kuinka karakteristisen funktion avulla on helppo osoittaa, että riippumattomien multinormaalisten (samanulotteisten) muuttujavektorien summa edelleen noudattaa multinormaalijakaumaa. Todistus perustuu tunnettuun karakteristisen funktion ominaisuuteen: Jos $\mathbf{X}$ ja $\mathbf{Y}$ ovat riippumattomia,

$$\phi_{\mathbf{X+Y}}(\mathbf{t}) = \phi_{\mathbf{X}}(\mathbf{t})\phi_{\mathbf{Y}}(\mathbf{t}) .$$

Olkoot siis muuttujavektorit $\mathbf{X}^{(1)}$, ..., $\mathbf{X}^{(N)}$ riippumattomia ja

$$\mathbf{X}^{(j)} \sim N(\mu^{(j)}, \Sigma^{(j)}),\ j = 1,2, ...,N .$$

Tällöin

$$\phi_{\mathbf{X}^{(1)} + \mathbf{X}^{(2)} + ... + \mathbf{X}^{(N)}}(\mathbf{t}) = \prod_{j=1}^{N} \exp(i\mathbf{t}'\mu^{(j)} - \tfrac{1}{2}\mathbf{t}'\Sigma^{(j)}\mathbf{t})$$

$$= \exp[\, i\mathbf{t}'(\mu^{(1)} + ... + \mu^{(N)}) - \tfrac{1}{2}\mathbf{t}'(\Sigma^{(1)} + ... + \Sigma^{(N)})\mathbf{t}]$$

eli saadusta karakteristisen funktion esitysmuodosta seuraa suoraan

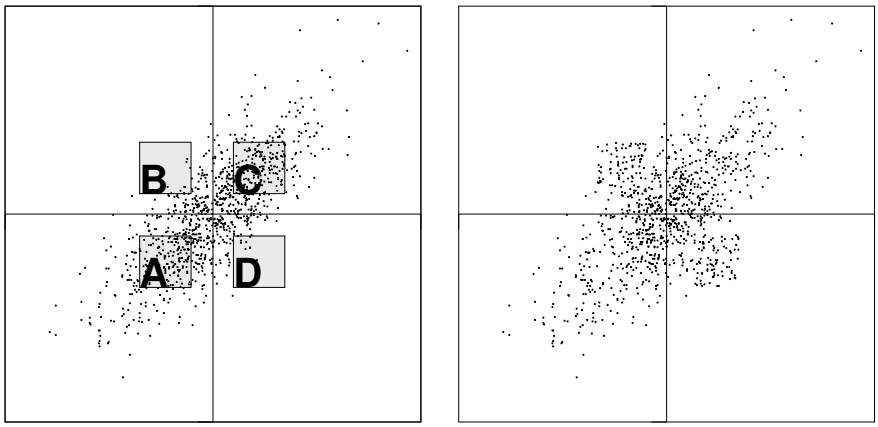
$$\mathbf{X}^{(1)} + ... + \mathbf{X}^{(N)} \sim N(\mu^{(1)} + ... + \mu^{(N)}, \Sigma^{(1)} + ... + \Sigma^{(N)}) .$$

2.2.7 Reunajakaumat ja multinormaalisuus

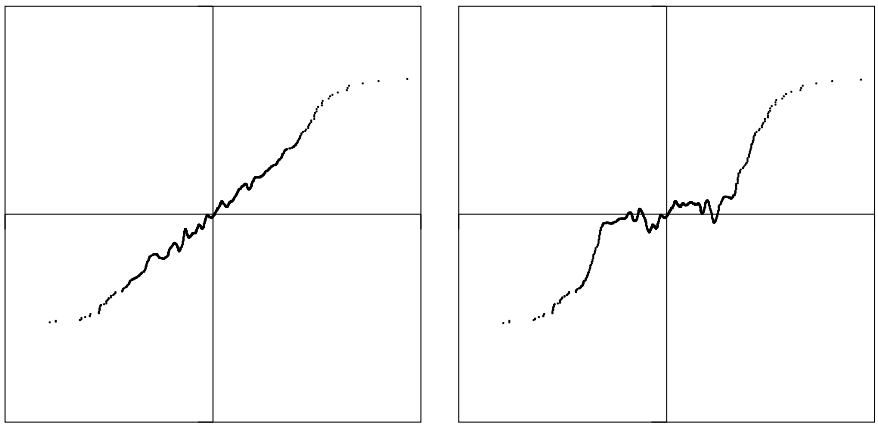
Usein varsinkin soveltajat kuvittelevat, että tutkittaessa usean muuttujan aineistoa, riittäisi multinormaalisuuden olemassaoloon reunajakaumien normaalisuus. Näin ei ole asian laita, vaan reunajakaumien ohella myös kaikkien mahdollisten muuttujien lineaaristen yhdistelmien tulee olla normaalisia.

Näytämme tässä pienellä simulointikokeella, että voi olla olemassa aineistoja, joissa reunajakaumat ovat normaalisia, mutta yhteisjakauma on kaukana multinormaalisesta.

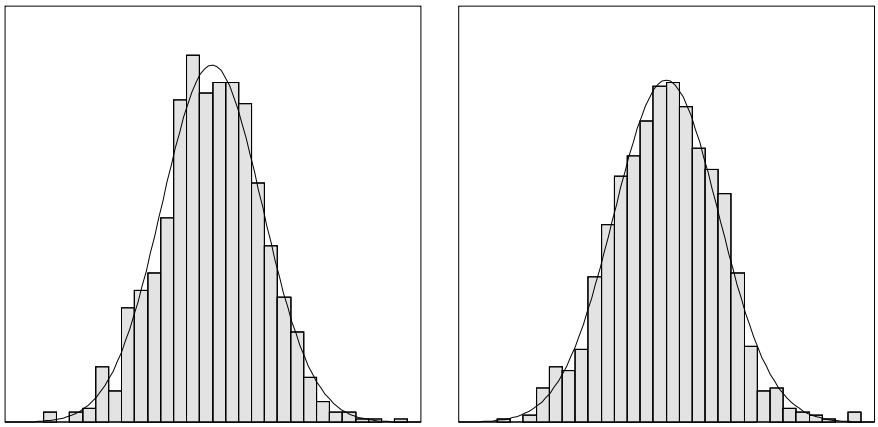
Seuraavassa kuvaparissa vasemmanpuoleinen esittää 1000 havainnon otosta 2-ulotteisesta *X*- ja *Y*-muuttujan normaalijakaumasta, jossa muuttujien keskiarvot ovat 0 ja hajonnat 1 sekä korrelaatiokerroin 0.8 . Oikeanpuoleinen kuva esittää samankokoista otosta, jossa todennäköisyydellä 0.5 sellainen havainto, joka osuu vasemmanpuoleisen kuvan A-ruutuun siirretäänkin 1.8 yksikköä ylöspäin B-ruutuun ja vastaavasti todennäköisyydellä 0.5 C-ruudun havainto siirtyy alaspäin D-ruutuun. Tämä muunnos ei muuta lainkaan *X*-arvoja ja symmetriasyistä se säilyttää *Y*-muuttujan jakauman, vaikka yksittäiset *Y*-arvot muuttuvatkin.



Kuitenkin jo silmämääräisesti on selvää, että näin muunnettu yhteisjakauma ei voi olla normaalinen. Tämä näkyy vielä paremmin piirtämällä regressiofunktioita approksimoivat regressiokäyrät (tässä Survon `SMOOTH`-operaatiolla), jolloin aidon multinormaalijakauman tapauksessa saadaan miltei suora viiva, mutta muunnetussa tapauksessa (oikeanpuoleinen kuva) mutkia syntyy melkoisesti.



Kuitenkin tarkasteltaessa muunnetun jakauman reunajakaumia, jotka on piirretty tässä histogrammeina, saadaan hyvin kauniisti normaalijakaumaa vastaavat tulokset. Verrattaessa X-muuttujan jakaumaa normaalijakaumaan χ^2-testillä, saadaan χ^2=18.77 vapausastein df=18, jolloin P=0.41. Vastaavat arvot Y-muuttujalle (kuvassa oikealla) ovat χ^2=20.52, df=19, P=0.36 .



3. Multinormaalinen otos

3.1 Parametrien estimointi

Tarkastelemme N riippumattoman havainnon

$$\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(N)}$$

otosta p-ulotteisesta multinormaalijakaumasta $N(\mu,\Sigma)$. Yleensä on syytä olettaa, että havaintojen lukumäärä N on huomattavasti suurempi kuin muuttujien lukumäärä p. Eräissä monimuuttujamenetelmissä ei ole oikeastaan mitään välttämättömiä rajoituksia; tulokset jäävät vain epäluotettavammiksi, kun havaintoluku on alhainen. Tässä yhteydessä on kuitenkin syytä olettaa, että $N>p$, sillä se esim. takaa todennäköisyydellä 1, että täysiasteisessa tapauksessa myös otoskovarianssi- ja korrelaatiomatriisi ovat säännöllisiä (täysiasteisia).

Koko otos (havaintoaineisto) voidaan kuvata $p\times N$-matriisina

$$\mathbf{X} = [\mathbf{x}^{(1)} \ \mathbf{x}^{(2)} \ \ldots \ \mathbf{x}^{(N)}] = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1N} \\ x_{21} & x_{22} & \ldots & x_{2N} \\ \ldots & \ldots & \ldots & \ldots \\ x_{p1} & x_{p2} & \ldots & x_{pN} \end{bmatrix}.$$

Multinormaalijakauman kannalta keskeisiä otossuureita ovat keskiarvovektori

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{x}^{(\alpha)} = \begin{bmatrix} \frac{1}{N}\sum_{\alpha=1}^{N} x_{1\alpha} \\ \ldots \\ \frac{1}{N}\sum_{\alpha=1}^{N} x_{p\alpha} \end{bmatrix} = \begin{bmatrix} \bar{x}_1 \\ \ldots \\ \bar{x}_p \end{bmatrix}$$

ja momenttimatriisi

$$\mathbf{A} = \sum_{\alpha=1}^{N}(\mathbf{x}^{(\alpha)} - \bar{\mathbf{x}})(\mathbf{x}^{(\alpha)} - \bar{\mathbf{x}})' = [a_{ij}],$$

jonka alkiot ovat

$$a_{ij} = \sum_{\alpha=1}^{N}(x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j), \quad i,j = 1,2, ...,p .$$

Matriisin $\mathbf{A}$ avulla määritellään edelleen otoskovarianssimatriisi

$$\mathbf{S} = \frac{1}{N\text{-}1}\mathbf{A} = [s_{ij}].$$

Matriisit **A** ja **S** ovat ei-negatiivisesti definiittejä. Tämä todetaan kirjoittamalla **A** muodossa

$$\mathbf{A} = (\mathbf{X} - \overline{\mathbf{X}})(\mathbf{X} - \overline{\mathbf{X}})' ,$$

missä

$$\overline{\mathbf{X}} = [\ \overline{\mathbf{x}}\ \overline{\mathbf{x}} \ldots \overline{\mathbf{x}}\ ]$$

on samaa muotoa kuin havaintomatriisi **X**, mutta jokainen havaintoarvo on korvattu ao. muuttujan keskiarvolla. Kun merkitään

$$\mathbf{C} = \mathbf{X} - \overline{\mathbf{X}} ,$$

niin neliömuoto

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{C}\mathbf{C}'\mathbf{x} = (\mathbf{C}'\mathbf{x})'(\mathbf{C}'\mathbf{x}) = \mathbf{y}'\mathbf{y} \geq 0 , \text{ missä } \mathbf{y}=\mathbf{C}'\mathbf{x} ,$$

eli $\mathbf{A}\geq\mathbf{0}$. Koska **S** on vakiotekijää vaille sama kuin **A**, myös $\mathbf{S}\geq\mathbf{0}$. Itse asiassa voidaan todistaa, että esim. **S** on positiivisesti definiitti ($\mathbf{S}>0$) todennäköisyydellä 1, jos Σ on täysiasteinen ja $N>p$.

Harjoitustehtäväksi jätetään sen osoittaminen, että **A** voidaan kirjoittaa myös muodossa

$$\mathbf{A} = \sum_{\alpha=1}^{N} \mathbf{x}^{(\alpha)}\mathbf{x}^{(\alpha)'} - N\overline{\mathbf{x}}\,\overline{\mathbf{x}}' .$$

Voidaan todistaa, että $\overline{\mathbf{x}}$ ja **A** ovat multinormaalijakauman $N(\mu,\Sigma)$ tyhjentäviä otossuureita ja että $\overline{\mathbf{x}}$ ja $\mathbf{A}/N$ ovat parametrien μ ja Σ suurimman uskottavuuden estimaattorit.

Todistuksen osalta viittaamme esim. teokseen T.W. Anderson: An Introduction to Multivariate Statistical Analysis (Wiley 1958), ss. 44 - 47. Todettakoon tässä kuitenkin, että maksimoitava uskottavuusfunktion logaritmi on suoraan multinormaalijakauman tiheysfunktion mukaisesti

$$\log L(\mu,\Sigma) = -\tfrac{1}{2}[pN\log(2\pi) + N\log|\Sigma| + \sum_{\alpha=1}^{N} (\mathbf{x}^{(\alpha)} - \mu)'\,\Sigma^{-1}(\mathbf{x}^{(\alpha)} - \mu)]$$

ja se voidaan saattaa muotoon

$$\log L(\mu,\Sigma) = -\tfrac{1}{2}[pN\log(2\pi) + N\log|\Sigma| + \ \mathrm{tr}(\Sigma^{-1}\mathbf{A}) + N(\overline{\mathbf{x}} - \mu)'\Sigma^{-1}(\overline{\mathbf{x}} - \mu)] .$$

Tästä esityksestä nähdään suoraan, että uskottavuusfunktio riippuu otoksesta vain parametrien $\overline{\mathbf{x}}$ ja **A** kautta. Ne ovat siis tyhjentäviä otossuureita. Lisäksi nähdään, että funktion maksimipisteessä $\mu=\overline{\mathbf{x}}$ eli $\overline{\mathbf{x}}$ on odotusarvovektorin suurimman uskottavuuden estimaattori.

Se, että funktio maksimoituu Σ:n suhteen, kun $\Sigma=\mathbf{A}/N$, on hankalampi todistaa. Viittaamme tältä osin em. Andersonin kirjaan.

Analogisesti yhden muuttujan tapauksen kanssa kovarianssimatriisin Σ estimaattorina käytetään tavallisesti kuitenkin matriisia $\mathbf{S}=\mathbf{A}/(N-1)$, koska tämän odotusarvo on Σ eli se on harhaton estimaattori, kuten jatkossa tullaan näyttä-

mään.

Vastaavasti korrelaatiomatriisin P suurimman uskottavuuden estimaattoriksi saadaan tavanomainen tulomomenttikorrelaatiokertoimista muodostuva matriisi $\mathbf{R} = [r_{ij}]$, missä

$$r_{ij} = \frac{s_{ij}}{s_i s_j}\,,\ i,j = 1,2,\ldots,p \quad \text{ja} \quad s_i^2 = s_{ii}\,.$$

Kun otetaan merkintä $\mathbf{D}_s$ hajontojen s_1, s_2, ..., s_p muodostamalle lävistäjämatriisille, saadaan yhteys

$$\mathbf{R} = \mathbf{D}_s^{-1}\mathbf{S}\mathbf{D}_s^{-1}\,.$$

On helppo näyttää, että $\mathbf{R}\geq 0$. Tilanteessa $\Sigma>0$ ja $N>p$ **R** on jopa positiivisesti definiitti ($\mathbf{R}>\mathbf{0}$) todennäköisyydellä 1.

Myös esim. osittaiskovarianssien, osittaiskorrelaatiokertoimien ja yhteiskorrelaatiokertoimien suurimman uskottavuuden estimaattorit saadaan vastinkaavoilla em. estimaattoreista $\bar{\mathbf{x}}$, **S** ja **R**.

3.2 Otossuureiden jakaumista

Yhden muuttujan normaalisen otoksen tapauksessa tiedetään, että otoskeskiarvo ja otosvarianssi ovat riippumattomia satunnaissuureita. Otoskeskiarvo noudattaa edelleen normaalijakaumaa alkuperäisellä odotusarvolla mutta pienemmällä hajonnalla ja otosvarianssin jakauma on vakiotekijää vaille χ^2-jakauma.

Multinormaalijakauman tapauksessa pätee vastaavien otossuureiden riippumattomuuden osalta sama tulos. Myös otoskeskiarvovektori on edelleen multinormaalinen ja otoskovarianssimatriisi noudattaa ns. Wishart-jakaumaa, joka on χ^2-jakauman moniulotteinen yleistys. Tulemme nyt johtamaan nämä tulokset.

Kuten yksiulotteisessakin tilanteessa, päättelyt perustuvat otoksen ortogonaaliseen muunnokseen, jolla erotetaan toisistaan keskiarvoja ja kovariansseja koskevat termit. Tämän vuoksi näytetään ensin toteen hiukan yleisempi apulause:

Olkoot $\mathbf{x}^{(1)}$, ..., $\mathbf{x}^{(N)}$ riippumattomia satunnaisvektoreita ja

$$\mathbf{x}^{(\alpha)} \sim N(\mu^{(\alpha)},\Sigma)\,,\ \alpha = 1,\ldots,N\,.$$

Olkoon edelleen $\mathbf{C}=[c_{\alpha\beta}]$ ortogonaalinen $N\times N$-matriisi. Merkitään

$$\nu^{(\alpha)} = \sum_{\beta=1}^{N} c_{\alpha\beta}\mu^{(\beta)}\,.$$

Tällöin on voimassa

$$\mathbf{y}^{(\alpha)} = \sum_{\beta=1}^{N} c_{\alpha\beta}\mathbf{x}^{(\beta)} \sim N(\nu^{(\alpha)},\Sigma)\,,\ \alpha = 1,\ldots,N$$

ja muuttujavektorit $\mathbf{y}^{(1)}$, ..., $\mathbf{y}^{(N)}$ ovat riippumattomia.

Todistukseksi riittää osoittaa, että **y**-muuttujien odotusarvot ja kovarianssimatriisit ovat väitteen mukaiset, sillä ne ovat ilman muuta multinormaalisesti jakautuneita riippumattomien **x**-muuttujien lineaarisina kombinaatioina.

Odotusarvojen osalta tilanne on selvä, sillä

$$E(\mathbf{y}^{(\alpha)}) = \sum_{\beta=1}^{N} c_{\alpha\beta} E(\mathbf{x}^{(\beta)}) = \nu^{(\alpha)} , \quad \alpha = 1, ..., N .$$

Tutkitaan sitten kahden **y**-vektorin $\mathbf{y}^{(\alpha)}$ ja $\mathbf{y}^{(\gamma)}$ välistä kovarianssimatriisia:

$$\begin{aligned}
&\mathrm{cov}(\mathbf{y}^{(\alpha)},\mathbf{y}^{(\gamma)}) = E[(\mathbf{y}^{(\alpha)} - \nu^{(\alpha)})(\mathbf{y}^{(\gamma)} - \nu^{(\gamma)})'] \\
&= E[\sum_{\beta=1}^{N} c_{\alpha\beta}(\mathbf{x}^{(\beta)} - \mu^{(\beta)})][\sum_{\varepsilon=1}^{N} c_{\gamma\varepsilon}(\mathbf{x}^{(\varepsilon)} - \mu^{(\varepsilon)})'] \\
&= \sum_{\beta=1}^{N}\sum_{\varepsilon=1}^{N} c_{\alpha\beta}c_{\gamma\varepsilon} E[(\mathbf{x}^{(\beta)} - \mu^{(\beta)})(\mathbf{x}^{(\varepsilon)} - \mu^{(\varepsilon)})'] \\
&= \sum_{\beta=1}^{N}\sum_{\varepsilon=1}^{N} c_{\alpha\beta}c_{\gamma\varepsilon}\delta_{\beta\varepsilon}\Sigma \quad (\delta_{\beta\varepsilon}=1, \text{ jos } \beta=\varepsilon, \text{ muuten } \delta_{\beta\varepsilon}=0) \\
&= \sum_{\beta=1}^{N} c_{\alpha\beta}c_{\gamma\beta}\Sigma = \delta_{\alpha\gamma}\Sigma
\end{aligned}$$

eli muuttujavektorien $\mathbf{y}^{(\alpha)}$ ja $\mathbf{y}^{(\gamma)}$ välinen kovarianssimatriisi on $\mathbf{0}$, kun $\alpha\neq\gamma$, mikä merkitsee samalla näiden muuttujavektoreiden riippumattomuutta. Jos taas $\alpha=\gamma$, kovarianssimatriisi on Σ, kuten väitettiin.

Toisena aputuloksena tarvitsemme seuraavan:

$$\sum_{\alpha=1}^{N} \mathbf{x}^{(\alpha)}\mathbf{x}^{(\alpha)'} = \sum_{\alpha=1}^{N} \mathbf{y}^{(\alpha)}\mathbf{y}^{(\alpha)'} .$$

Tämä todetaan oikeaksi suoralla laskulla

$$\begin{aligned}
&\sum_{\alpha=1}^{N} \mathbf{y}^{(\alpha)}\mathbf{y}^{(\alpha)'} = \sum_{\alpha=1}^{N} (\sum_{\beta=1}^{N} c_{\alpha\beta}\mathbf{x}^{(\beta)})(\sum_{\gamma=1}^{N} c_{\alpha\gamma}\mathbf{x}^{(\gamma)'}) \\
&= \sum_{\beta=1}^{N}\sum_{\gamma=1}^{N} (\sum_{\alpha=1}^{N} c_{\alpha\beta}c_{\alpha\gamma})\mathbf{x}^{(\beta)}\mathbf{x}^{(\gamma)'} = \sum_{\beta=1}^{N}\sum_{\gamma=1}^{N} \delta_{\beta\gamma}\mathbf{x}^{(\beta)}\mathbf{x}^{(\gamma)'} = \sum_{\beta=1}^{N} \mathbf{x}^{(\beta)}\mathbf{x}^{(\beta)'} .
\end{aligned}$$

Sovellamme nyt näitä aputuloksia multinormaalijakaumasta $N(\mu,\Sigma)$ saatuun N havainnon otokseen $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, ..., $\mathbf{x}^{(N)}$. Valitsemme $N\times N$ ortogonaalisen matriisin **C** siten, että sen viimeisen vaakarivin jokainen alkio on $1/\sqrt{N}$.
Otoksesta lasketun momenttimatriisin **A** voimme kirjoittaa muodossa

$$\mathbf{A} = \sum_{\alpha=1}^{N} (\mathbf{x}^{(\alpha)} - \bar{\mathbf{x}})(\mathbf{x}^{(\alpha)} - \bar{\mathbf{x}})' = \sum_{\alpha=1}^{N} \mathbf{x}^{(\alpha)}\mathbf{x}^{(\alpha)'} - N\bar{\mathbf{x}}\,\bar{\mathbf{x}}' .$$

Olkoon nyt

$$\mathbf{z}^{(\alpha)} = \sum_{\beta=1}^{N} c_{\alpha\beta}\mathbf{x}^{(\beta)} , \ \alpha = 1,2, ...,N ,$$

jolloin erityisesti viimeinen näistä on

$$\mathbf{z}^{(N)} = \sqrt{N}\,\bar{\mathbf{x}} ,$$

koska matriisin **C** viimeisen vaakarivin jokainen alkio on $1/\sqrt{N}$. Käyttämällä hyväksi tulosta

$$\sum_{\alpha=1}^{N} \mathbf{z}^{(\alpha)}\mathbf{z}^{(\alpha)'} = \sum_{\alpha=1}^{N} \mathbf{x}^{(\alpha)}\mathbf{x}^{(\alpha)'} ,$$

toteamme, että

$$\mathbf{A} = \sum_{\alpha=1}^{N} \mathbf{x}^{(\alpha)}\mathbf{x}^{(\alpha)'} - N\bar{\mathbf{x}}\,\bar{\mathbf{x}}' = \sum_{\alpha=1}^{N} \mathbf{z}^{(\alpha)}\mathbf{z}^{(\alpha)'} - \mathbf{z}^{(N)}\mathbf{z}^{(N)'} = \sum_{\alpha=1}^{N-1} \mathbf{z}^{(\alpha)}\mathbf{z}^{(\alpha)'} .$$

Koska muuttujavektorit

$\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, ..., \mathbf{z}^{(N)}$

apulauseen perusteella ovat riippumattomia satunnaissuureita ja otoskeskiarvo $\bar{\mathbf{x}}$ riippuu vain niistä viimeisestä sekä momenttimatriisi **A** N-1 ensimmäisestä, voimme todeta, että $\bar{\mathbf{x}}$ ja **A** ovat toisistaan riippumattomia.

Edelleen apulauseen perusteella

$$\mathbf{z}^{(\alpha)} \sim N(\sum_{\beta=1}^{N} c_{\alpha\beta}\mu, \Sigma) , \ \alpha = 1,2, ...,N .$$

Tällöin erityisesti

$$\mathbf{z}^{(N)} \sim N(\sqrt{N}\mu, \Sigma)$$

eli

$$\bar{\mathbf{x}} = \mathbf{z}^{(N)}/\sqrt{N} \sim N(\mu, \Sigma/N) .$$

Kun $\alpha \neq N$,

$$\mathrm{E}(\mathbf{z}^{(\alpha)}) = \sum_{\beta=1}^{N} c_{\alpha\beta}\mu = \mu \sum_{\beta=1}^{N} c_{\alpha\beta} = 0 ,$$

sillä koska ortogonaalisen matriisin **C** viimeisen vaakarivin alkiot ovat samoja, kaikkien muiden vaakarivien alkioiden summat ortogonaalisuudesta johtuen ovat nollia.

Yhteenvetona voimme todeta, että multinormaalisesta otoksesta laskettu otoskeskiarvovektori ja momenttimatriisi ovat riippumattomia satunnaissuureita. Otoskeskiarvovektori noudattaa multinormaalijakaumaa alkuperäisellä odotusarvolla μ, mutta kovarianssimatriisi tulee jaetuksi otoksen koolla N.

Momenttimatriisi **A** on jakautunut kuten summa

$$\sum_{\alpha=1}^{N-1} \mathbf{z}^{(\alpha)}\mathbf{z}^{(\alpha)'} ,$$

missä satunnaisvektorit $\mathbf{z}^{(1)}$, $\mathbf{z}^{(2)}$, ..., $\mathbf{z}^{(N-1)}$ ovat riippumattomia ja niistä jokainen noudattaa multinormaalijakaumaa $N(\mathbf{0},\Sigma)$.

Tätä jakaumaa, joka riippuu parametreista N-1 ja Σ, sanotaan *Wishart*-jakaumaksi ja merkitään $\mathbf{A} \sim W(N\text{-}1,\Sigma)$. Näemme välittömästi, että

$$E(\mathbf{A}) = \sum_{\alpha=1}^{N-1} E(\mathbf{z}^{(\alpha)}\mathbf{z}^{(\alpha)'}) = \sum_{\alpha=1}^{N-1} \mathrm{cov}(\mathbf{z}^{(\alpha)}) = (N\text{-}1)\Sigma .$$

Täten siis otoskovarianssimatriisi $\mathbf{S}=\mathbf{A}/(N\text{-}1)$ on kovarianssimatriisin Σ harhaton estimaattori multinormaalijakaumassa.

Samoin kuin multinormaalijakauma on tavallisen yksiulotteisen normaalijakauman yleistys, Wishart-jakauma, jonka on esitellyt *John Wishart* v. 1928, on χ^2-jakauman yleistys. Jätämme harjoitustehtäväksi todeta, että erikoistapauksessa p=1 edellä johdetut tulokset palautuvat tuttuihin normaalista otosta koskeviin tuloksiin ja erityisesti $W(n,1)$-jakauma on sama kuin χ^2-jakauma n vapausasteella.

Huomattakoon kuitenkin, että p-ulotteisessa tilanteessa Wishart-jakauman todellinen ulotteisuusluku on $p(p+1)/2$ eli tämän jakauman hallitseminen on hankalampaa, kuin sen taustana olevan multinormaalijakauman.

Tarkempaa tietoutta Wishart-jakaumasta löytyy mm. kirjoista *T.W.Anderson* (1958), *C.R.Rao* (1965) ja *G.A.F.Seber* (1984). Tulemme käyttämään näissä kirjoissa esitettyjä tuloksia esim. multinormaalijakaumaan liittyvissä tilastollisissa testeissä.

3.3 Multinormaalisen otoksen simulointi

Kun jatkossa esittelemme erilaisia multinormaalijakaumaan perustuvia menetelmiä, tulemme aitojen aineistojen ohella käyttämään myös simuloituja, Monte Carlo-menetelmällä tehtyjä aineistoja. Kokeillessamme jotain menetelmää aitoon aineistoon ja havaitessamme, ettei menetelmä anna toivottuja tuloksia, emme voi olla varmoja siitä, johtuuko epäonnistuminen menetelmän huonoudesta vai siitä, ettei aineisto ole otos multinormaalijakaumasta. Simuloitujen aineistojen kohdalla jälkimmäiselle epäilylle ei ole sijaa ja niinpä ne tarjoavat hyvät edellytykset eri menetelmien kokeiluun ja käyttökelpoisuuden arviointiin.

Multinormaalisen otoksen simulointi tapahtuu helpoimmin suoraan konstruktiivisen määritelmän (2) $\mathbf{X}=\mathbf{AV}+\mu$ avulla. Jos siis tarvitsemme havaintoja jakaumasta $N(\mu,\Sigma)$, laskemme kovarianssimatriisin Σ spektraalihajotelman $\Sigma=\mathbf{U}\Lambda\mathbf{U}'$ ja valitsemme $\mathbf{A}=\mathbf{U}\Lambda^{1/2}$, jolloin $\Sigma=\mathbf{AA}'$.

Survossa tätä varten on käytettävissä sukro MNSIMUL, joka luo N havainnon otoksen jakaumasta $N(\mu,\Sigma)$. Parametrit annetaan kahden matriisitiedoston **R** ja **M** avulla, missä **R**:n tulee vastata rakenteeltaan Survon CORR-operaatiolla saatua CORR.M-tiedostoa (korrelaatiomatriisi) ja **M**:n CORR-operaatiolla saatua MSN.M-tiedostoa, jonka kahtena ensimmäisenä pystyrivinä ovat odotusarvot ja keskihajonnat.

```
 16  1 SURVO 84C EDITOR Sat Feb 12 14:26:48 1994        D:\M\MONI\ 100 100 0
  1 *
  2 *MATRIX R
  3 *///  X1   X2   X3
  4 *X1   1    0.4 -0.7
  5 *X2   0.4  1   -0.2
  6 *X3  -0.7 -0.2  1
  7 *
  8 *MATRIX M
  9 *///  Keski  Hajonta
 10 *X1    0       1
 11 *X2    1       2
 12 *X3   -2       0.5
 13 *
 14 *MAT SAVE R
 15 *MAT SAVE M
 16 */MNSIMUL R,M,OTOS,1000  /  RND=rand(1)
 17 *
 18 *CORR OTOS,CUR+1_
 19 *Means, std.devs and correlations of OTOS  N=1000
 20 *Variable  Mean        Std.dev.
 21 *X1       -0.016924    1.030074
 22 *X2        0.999173    2.011132
 23 *X3       -1.982124    0.484458
 24 *Correlations:
 25 *             X1      X2      X3
 26 * X1           1.0000  0.3699 -0.6876
 27 * X2           0.3699  1.0000 -0.1868
 28 * X3          -0.6876 -0.1868  1.0000
 29 *
```

Oheinen esimerkki näyttää 3 muuttujan tapauksessa, miten MNSIMUL-sukroa

käytetään. Lähtömatriisit on kirjoitettu riveille 2-12 ja ne talletetaan matriisitiedostoiksi R.MAT ja M.MAT riveillä 14 ja 15 olevilla MAT SAVE-komennoilla. Rivin 16 /MNSIMUL-sukrokomento generoi matriisien **R** ja **M** avulla havaintotiedoston OTOS, johon lasketaan 1000 havaintoa.

Tulos on tarkastettu rivin 18 CORR-komennolla, jonka antamat tulokset ovat riveillä 19-28. Nähdään välittömästi, että estimoidut keskiarvot, hajonnat ja korrelaatiokertoimet näyttävät riittävän hyvin vastaavan jakauman teoreettisia parametrin arvoja.

Sukro /MNSIMUL käyttää RND-täsmennyksellä määriteltyä generaattoria luodessaan tasaisesti jakautuneita pseudosatunnaislukuja, jotka muunnetaan muuttujia V vastaaviksi riippumattomiksi $N(0,1)$-satunnaisluvuiksi. Tässä on generaattoriksi valittu rand(1) rivillä 16.

Jos RND-täsmennystä ei anneta, käytetään funktiota rnd(0) eli koneen kellosta riippuvaa siemenlukua, jolloin koetta toistettaessa saadaan joka kerralla eri tulokset. Niiden tulisi kuitenkin lähes aina vastata odotettuja arvoja etenkin silloin, kun otoskoko (tässä 1000) on riittävän suuri.

Jos muuttujia on vain kaksi, suorempi generointitapa on luoda ensin kaksi riippumatonta satunnaisarvoa V_1 ja V_2 jakaumasta $N(0,1)$ ja laskea lopulliset muuttujat X_1 ja X_2 kaavoilla

$$X_1 = \sigma_1 V_1 + \mu_1$$

$$X_2 = \sigma_2(\rho V_1 + \sqrt{1-\rho^2}\, V_2) + \mu_2 \,.$$

Se että näin syntyy muuttujapari (X_1,X_2), jonka odotusarvot ovat (μ_1,μ_2), hajonnat (σ_1,σ_2) ja korrelaatiokerroin ρ, jätetään harjoitustehtäväksi.

Seuraava Survon laskentakaavio osoittaa, miten nämä kaavat toimivat käytännössä:

```
 32  1 SURVO 84C EDITOR Fri Feb 11 16:07:45 1994        D:\M\MONI\ 100 100 0
 31 *
 32 *   Pituuden ja painon arvonta:
 33 *              keskiarvo      hajonta
 34 *   Pituus      m1=175 cm      s1=6
 35 *   Paino       m2=72 kg       s2=5
 36 *   Pituuden ja painon korrelaatiokerroin r=0.82
 37 *
 38 *   V1=probit(rnd(0)) V2=probit(rnd(0))
 39 *   Pituus=int(s1*V1+m1)          int() ottaa lausekkeen kokonaisosan
 40 *    Paino=int(s2*(r*V1+sqrt(1-r*r)*V2)+m2)
 41 *
 42 *   Pituus.=169           Paino.=69
 43 *
```

Tässä kaaviossa simuloidaan "ihmispopulaation" käyttäytymistä pituuden ja painon suhteen. Rivillä 38 rnd(0) tarkoittaa tasaisesti väliltä (0,1) arvottua satunnaislukua ja probit-funktio (normaalijakauman kertymäfunktion käänteisfunktio) muuntaa sen (0,1)-normaaliseksi. Varsinaiset laskukaavat ovat riveillä 39-40 ja aktivoimalla kumpi tahansa rivin 42 kohteista saadaan tälle riville aina uusia pituuden ja painon arvoja riveillä 34-36 annettujen perusparametrien ja 2-ulotteisen normaalijakauman mukaisesti.

Tämä laskentakaavio on helppo ottaa pohjaksi, jos halutaan tallentaa ko. jakaumaa noudattava otos havaintotiedostoon tai -taulukkoon Survon VAR-operaatiolla. Seuraava Survo-kaavio näyttää, miten 30 havainnon otos luodaan. Tässä oletetaan, että kaavio on suoraa jatkoa edellisen kaavion riveille 31-43:

```
 26  1 SURVO 84C EDITOR Fri Feb 11 16:08:30 1994          D:\M\MONI\ 100 100 0
 43 *
 44 *VAR Pituus,Paino TO OTOS2  / Aktivoimalla uudelleen syntyy uusia otoksia
 45 *DATA OTOS2,A,A+29,N,M
 46 M   111     111
 47 N  Pituus  Paino
 48 A   180      73
 49 *   182      77
 50 *   174      70
 51 *   177      76
 52 *   175      68
 53 *   171      68
 54 *   178      71
 55 *   173      70
 56 *   172      71
 57 *   172      66
 58 *   188      80
 59 *   174      64
 60 *   179      77
 61 *   170      72
 62 *   179      75
 63 *   167      65
 64 *   173      69
 65 *   178      71
 66 *   174      67
 67 *   170      68
 68 *   179      78
 69 *   172      69
 70 *   176      72
 71 *   170      69
 72 *   180      75
 73 *   173      74
 74 *   184      75
 75 *   166      64
 76 *   175      68
 77 *   187      80
 78 *
 79 *CORR OTOS2,CUR+1
 80 *Means, std.devs and correlations of OTOS2  N=30
 81 *Variable  Mean        Std.dev.
 82 *Pituus    175.6000    5.353697
 83 *Paino     71.40000    4.515109
 84 *Correlations:
 85 *             Pituus  Paino
 86 * Pituus       1.0000  0.8285
 87 * Paino        0.8285  1.0000
 88 *
```

Tulos on tarkastettu laskemalla otoksesta saadut tunnusluvut CORR-operaatiolla.

Vielä välittömämmin yleistä kaksiulotteista normaalijakaumaa luodaan kaavoilla

$$X_1 = \mu_1 + \sigma_1 \sqrt{-2\log(U_2)}\,\cos(2\pi U_1)\,,$$

$$X_2 = \mu_2 + \sigma_2 \sqrt{-2\log(U_2)}\,\sin(2\pi U_1 + \arcsin(\rho))\,,$$

missä U_1 ja U_2 ovat riippumattomia, tasaisesti välillä (0,1) jakautuneita satunnaislukuja. Erikoistapauksessa $\rho=0$ (ja $\mu_1=\mu_2=0$, $\sigma_1=\sigma_2=1$), jolloin saadaan kaksi riippumatonta $N(0,1)$-muuttujaa, kaavat tunnetaan *Box-Müller*in

nimellä. Nyt esitetyn yleistyksen havaitsin aikoinaan johtaessani hajontaellipsien yhtälöt, jotka mainitaan kohdassa 4.2.1 .

Viimeksi mainituista kaavoista on hyötyä kaksiulotteisen normaalijakauman generoinnissa, jos ne ohjelmoidaan suoraan esim. C-kielellä. Survossa kaikki toimituskenttään kirjoitettujen kaavojen mukaiset laskennat tapahtuvat kuitenkin tulkkaamalla, jolloin laskentanopeus riippuu enemmän kaavojen pituudesta kuin niiden matemaattisesta yksinkertaisuudesta. Tämän vuoksi aikaisemmin todettu tapa on Survossa nopeampi.

3.4 Multinormaalijakaumaan liittyviä testejä

Multinormaalijakauman tapauksessa voidaan tutkia hyvin monenlaisia hypoteeseja. Keskitytään ensin odotusarvoja koskeviin testeihin ja tarkastellaan tapausta, jossa nollahypoteesina on $\mu=\mu^{(0)}$, kun oletetaan kovarianssimatriisi Σ tunnetuksi. Tarvitsemme seuraavan apulauseen:

Jos $\mathbf{Y}$ noudattaa p-ulotteista normaalijakaumaa $N(\mathbf{0},\Sigma)$, niin neliömuoto $\mathbf{Y}'\Sigma^{-1}\mathbf{Y}$ noudattaa χ^2-jakaumaa p vapausasteella.

Multinormaalijakauman konstruktiivisen määritelmän (2) mukaisesti $\mathbf{Y}$ voidaan lausua muodossa

$$\mathbf{Y} = \mathbf{AV} ,$$

missä $\mathbf{V} = (V_1,V_2,...,V_p) \sim N(\mathbf{0},\mathbf{I})$. Matriisi $\mathbf{A}$ on säännöllinen ja $\mathbf{AA}'=\Sigma$. Tällöin

$$\mathbf{Y}'\Sigma^{-1}\mathbf{Y} = \mathbf{V}'\mathbf{A}'(\mathbf{AA}')^{-1}\mathbf{AV} = \mathbf{V}'\mathbf{V} = V_1^2 + V_2^2 + ... + V_p^2$$

eli riippumattomien (0,1)-normaalisten muuttujien $V_1,V_2,...,V_p$ neliöiden summana $\mathbf{Y}'\Sigma^{-1}\mathbf{Y}$ noudattaa χ^2-jakaumaa p vapausasteella.

Koska multinormaalisessa otoksessa

$$\sqrt{N}\,(\bar{\mathbf{x}} - \mu) \sim N(\mathbf{0},\Sigma) ,$$

apulauseen mukaan

$$N(\bar{x} - \mu)'\,\Sigma^{-1}(\bar{x} - \mu) \sim \chi_p^2 .$$

Näin ollen hypoteesin H_0: $\mu=\mu^{(0)}$ ollessa voimassa

$$N(\bar{x} - \mu^{(0)})'\,\Sigma^{-1}(\bar{x} - \mu^{(0)}) \sim \chi_p^2 .$$

Valitaan kriittinen taso ε ja olkoon $P\{\chi_p^2 \geq \chi_p^2(\varepsilon)\} = \varepsilon$.

Testattaessa hypoteesia H_0: $\mu=\mu^{(0)}$ hypoteesia H_1: $\mu\neq\mu^{(0)}$ vastaan, testin kriittinen alue on siis

$$N(\bar{x} - \mu^{(0)})'\,\Sigma^{-1}(\bar{x} - \mu^{(0)}) \geq \chi_p^2(\varepsilon) .$$

Tämä testi voidaan johtaa myös osamääräperiaatteella, mikä jätetään harjoi-

tustehtäväksi.

Käytännössä kovarianssimatriisia Σ ei yleensä tunneta, joten testi tässä muodossa on käyttökelpoinen vain hyvin suurilla otoskoilla, kun Σ korvataan otoskovarianssimatriisilla **S**. Parempi on kuitenkin käyttää Hotellingin T^2-testiä, joka on täysin analoginen, mutta jossa testisuureen jakauma nollahypoteesin tapauksessa muuntuu F-jakaumaksi.

3.4.1 Mahalanobis-etäisyydet

Yksittäisen havainnon **x** poikkeamalle jakauman $N(\mu,\Sigma)$ keskipisteestä, kun μ ja Σ korvataan otoksesta lasketuilla estimaateillaan, on edellisen tarkastelun valossa sopiva käyttää mittaa

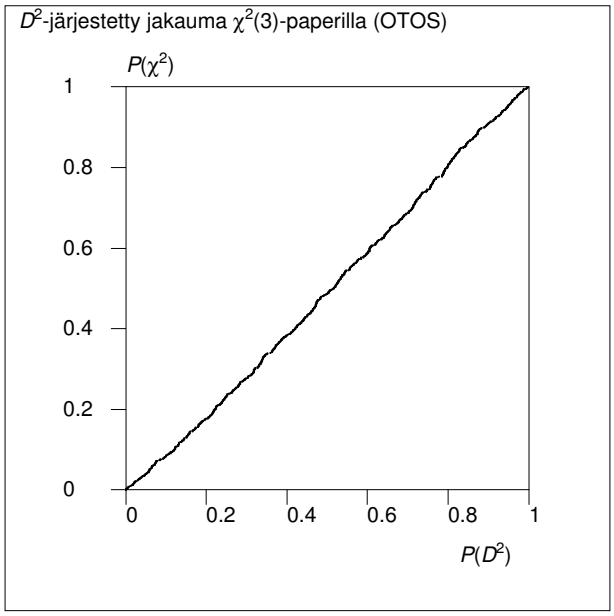
$$D^2 = (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) ,$$

jota sanotaan *Mahalanobis*-etäisyydeksi. Jos **S** olisi **I**, kyseessä on tavallisen euklidisen etäisyyden neliö. D^2 on euklidista etäisyyttä parempi mitta, koska se ottaa huomioon muuttujien keskinäisen riippuvuuden eikä ole riippuvainen käytetyistä mitta-asteikoista.

Mahalanobis-etäisyydet tarjoavat erään mahdollisuuden tutkia otoksen multinormaalisuutta, sillä suurilla otoskoilla, edellä todetun perusteella, D^2 noudattaa χ^2-jakaumaa p vapausasteella.

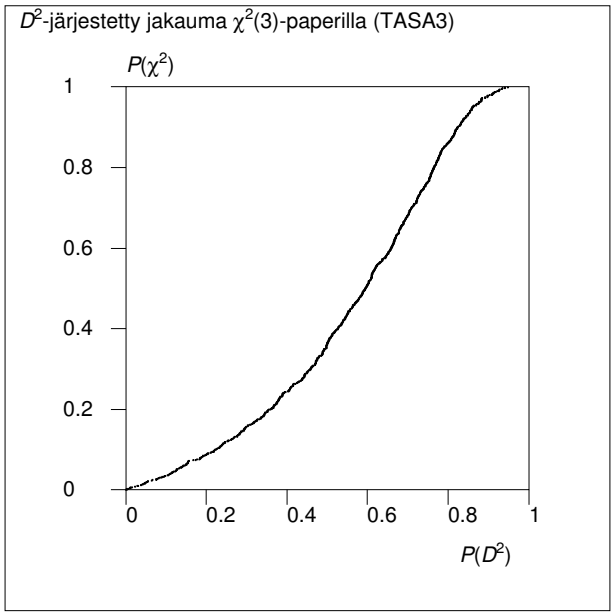
Esimerkkinä tarkastelemme edellä luvussa *Multinormaalisen otoksen simulointi* luotua 3 muuttujan multinormaalista otosta. Seuraava Survo-kaavio näyttää, miten ko. otoksesta lasketaan Mahalanobis-etäisyyksien muunnokset välin (0,1) tasaiseen jakaumaan (MAHAL-operaatio) ja tämän perusteella piirretään järjestetty D^2-arvojen otos $\chi^2(3)$-paperille. Tällöin multinormaalisen otoksen tulisi kuvautua likimain suoralle $y=x$.

```
 17  1 SURVO 84C EDITOR Tue Feb 15 09:12:01 1994        D:\M\MONI\ 100 100 0
  1 *
  2 *VAR C2=MISSING TO OTOS
  3 *MAHAL OTOS / VARS=X1(A),X2(A),X3(A),C2(P)
  4 *FILE SORT OTOS BY C2 TO OTOS2
  5 *VAR P=(ORDER-0.5)/N TO OTOS2
  6 *GPLOT OTOS2,C2,P_/ SCALE=0(0.2)1 POINT=11
  7 *
```



Jos samaa yritetään tehdä esim. 3 muuttujan 1000 havainnon otoksella, jossa muuttujat ovat tasaisesti välillä (0,1) jakautuneita, otos (TASA3) luodaan ja piirretään $\chi^2(3)$-paperille esim. seuraavasti:

```
 18  1 SURVO 84C EDITOR Tue Feb 15 09:15:58 1994        D:\M\MONI\ 100 100 0
 10 *
 11 *FILE CREATE TASA3,20,5
 12 *FIELDS:
 13 *1 N 4 U1
 14 *2 N 4 U2
 15 *3 N 4 U3
 16 *END
 17 *
 18 *FILE INIT TASA3,1000
 19 *VAR U1,U2,U3 TO TASA3
 20 *U1=rand(1) U2=rand(1) U3=rand(1)
 21 *
 22 *VAR C2=MISSING TO TASA3
 23 *MAHAL TASA3 / VARS=U1(A),U2(A),U3(A),C2(P)
 24 *FILE SORT TASA3 BY C2 TO TASA32
 25 *VAR P=(ORDER-0.5)/N TO TASA32
 26 *GPLOT TASA32,C2,P_/ SCALE=0(0.2)1 POINT=11
 27 *
```



Suora muuttuu S:n muotoiseksi käyräksi, mikä osoittaa, ettei voi olla kyse multinormaalisesta otoksesta.

3.4.2 Hotellingin T^2-testi (yhden otoksen tapaus)

Tutkimme edelleen hypoteesin H_0: $\mu=\mu^{(0)}$ testaamista multinormaalisen otoksen tapauksessa, mutta nyt oletamme, ettei kovarianssimatriisia tunneta. Tämä tilanne on p-ulotteinen yleistys tavallisesta yhden otoksen t-testistä ja se voidaankin johtaa tämän perusteella eräänlaisella maksimointiperiaatteella.

Hypoteesi H_0 on sama kuin hypoteesi:

$\mathbf{a}'\mu = \mathbf{a}'\mu^{(0)}$ on voimassa kaikilla vektoreilla $\mathbf{a}=(a_1,a_2,...,a_p)$.

Jokaisella vektorilla $\mathbf{a}$

$$t(\mathbf{a}) = \sqrt{N}\ \frac{(\bar{\mathbf{x}} - \mu^{(0)})'\mathbf{a}}{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}}$$

on hypoteesin $H_0(\mathbf{a})$: $\mathbf{a}'\mu = \mathbf{a}'\mu^{(0)}$ tavanomainen t-testisuure. Pyrimme nyt määräämään sen vektorin $\mathbf{a}$, joka maksimoi tämän testisuureen itseisarvon tai sen neliön, mikä on teknisesti yksinkertaisempaa. Etsimme siis p-ulotteisesta otosavaruudesta sen suunnan $\mathbf{a}$, jossa tavallisen t-testin $H_0(\mathbf{a})$-hypoteesi olisi heikoimmin voimassa.

Maksimointitehtävä (hankalan nimittäjän välttämiseksi) on paras pukea muotoon: On maksimoitava

$[\sqrt{N}\,[(\bar{\mathbf{x}} - \mu^{(0)})'\mathbf{a}]^2$ ehdolla $\mathbf{a}'\mathbf{S}\mathbf{a} =$ vakio .

Ottamalla käyttöön kovarianssimatriisin **S** Cholesky-hajotelman $\mathbf{S}=\mathbf{C}'\mathbf{C}$ ja määrittelemällä $\mathbf{b}=\mathbf{C}\mathbf{a}$, jolloin $\mathbf{a}=\mathbf{C}^{-1}\mathbf{b}$, tehtävä muuntuu muotoon: On maksimoitava

$[\sqrt{N}\,(\bar{\mathbf{x}} - \mu^{(0)})'\mathbf{C}^{-1}\mathbf{b}]^2$ ehdolla $\mathbf{b}'\mathbf{b} = \|\mathbf{b}\|^2 =$ vakio .

Merkitsemällä

$$\mathbf{u}' = \sqrt{N}\,(\bar{\mathbf{x}} - \mu^{(0)})'\mathbf{C}^{-1}$$

voimme kirjoittaa maksimoitavan lausekkeen muodossa $(\mathbf{u}'\mathbf{b})^2=\mathbf{b}'\mathbf{u}\mathbf{u}'\mathbf{b}$, joka saavuttaa maksiminsa ehdolla $\|\mathbf{b}\|=$vakio, kun **b** on matriisin $\mathbf{u}\mathbf{u}'$ suurinta ominaisarvoa vastaava ominaisvektori. Tämä $p\times p$-matriisi $\mathbf{u}\mathbf{u}'$ on vain astetta 1 ja sen ainoaa nollasta eroavaa ominaisarvoa vastaava ominaisvektori on **u**. Siis maksimin antava **b** on **u** ja maksimiarvo on

$$\mathbf{b}'\mathbf{u}\mathbf{u}'\mathbf{b}/\mathbf{b}'\mathbf{b} = \mathbf{u}'\mathbf{u} = N(\bar{\mathbf{x}} - \mu^{(0)})'\mathbf{C}^{-1}(\mathbf{C}^{-1})'(\bar{\mathbf{x}} - \mu^{(0)}) = N(\bar{\mathbf{x}} - \mu^{(0)})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \mu^{(0)}) .$$

Saatua testisuuretta merkitään

$$T^2 = N(\bar{\mathbf{x}} - \mu^{(0)})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \mu^{(0)})$$

ja voidaan osoittaa, että nollahypoteesin tapauksessa T^2 on kerroinvakiota vaille F-jakautunut eli

$$\frac{(N-p)}{(N-1)p}\,T^2 \sim F_{p,N-p}$$ (kts. esim. Anderson ss. 105-107).

T^2-testi voidaan johtaa myös osamääräperiaatteella (Anderson, luku 5).

Esim. Vertailu t-testiin kahden muuttujan X_1 ja X_2 tapauksessa:

Osoitamme nyt tapauksessa p=2, miten T^2-testisuure voidaan lausua yhden muuttujan t-testisuureiden lausekkeena.

Otoskovarianssimatriisi **S** ja sen käänteismatriisi $\mathbf{S}^{-1}$ ovat

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_1 s_2 r \\ s_1 s_2 r & s_2^2 \end{bmatrix} \qquad \mathbf{S}^{-1} = \frac{1}{s_1^2 s_2^2 (1-r^2)} \begin{bmatrix} s_2^2 & -s_1 s_2 r \\ -s_1 s_2 r & s_1^2 \end{bmatrix},$$

jolloin T^2 voidaan kirjoittaa muodossa

$$T^2 = \frac{N}{s_1^2 s_2^2 (1-r^2)} \times [\bar{x}_1-\mu_1 \ , \ \bar{x}_2-\mu_2] \begin{bmatrix} s_2^2 & -s_1 s_2 r \\ -s_1 s_2 r & s_1^2 \end{bmatrix} \begin{bmatrix} \bar{x}_1-\mu_1 \\ \bar{x}_2-\mu_2 \end{bmatrix}$$

$$= \frac{N}{s_1^2 s_2^2 (1-r^2)} [s_2^2(\bar{x}_1-\mu_1)^2 + s_1^2(\bar{x}_2-\mu_2)^2 - 2r(\bar{x}_1-\mu_1)(\bar{x}_2-\mu_2)]$$

$$= \frac{1}{1-r^2} (t_1^2 + t_2^2 - 2 r t_1 t_2) \ ,$$

missä

$$t_i = \sqrt{N} \frac{\bar{x}_i - \mu_i}{s_i} \ , \ i=1,2$$

ovat muuttujien X_1 ja X_2 tavallisia t-testisuureita.

Esityksestä

$$T^2 = \frac{1}{1-r^2} (t_1^2 + t_2^2 - 2 r t_1 t_2)$$

näemme selvästi, miten korrelaatiokerroin r vaikuttaa testisuureen T^2 arvoon.

Esim. jos $t_1>0$ ja $t_2>0$ mutta $r<0$, T^2-testi tällaisessa ristiriitatilanteessa voi hylätä nollahypoteesin, vaikka kumpikaan yksittäisistä t-testeistä ei sitä tekisikään. Jos kuitenkin myös $r>0$, T^2-testi muuttuu paljon konservatiivisemmaksi.

Tämä näkyy selvästi seuraavasta esimerkistä, jossa on oletettu, että N=100, t_1=t_2=1 ja jossa kummallakin t-testillä P=0.32 . Jos r=-0.7, kuitenkin T^2-testi antaa melkein merkitsevän eron (P=0.041). Jos sen sijaan r=0.7, ero nollahypoteesin mukaiseen tilanteeseen osoittautuu erillisiä t-testejä heikommaksi (P=0.56).

Alla olevan Survon laskentakaavion avulla on helppo tehdä lisää vastaavia vertailuja:

```
 61  1 SURVO 84C EDITOR Wed Feb 09 09:37:16 1994        D:\M\MONI\ 100 100 0
  1 *
  2 *   t-testin ja T2-testin vertailu 2 muuttujan tapauksessa
  3 *
  4 *   N=100  p=2                          Tulostustarkkuus:  ACCURACY=4
  5 *
  6 *   t1=1 t2=1  (oletetut t-testisuureen arvot)
  7 *
  8 *Hylkäystodennäköisyys kummallakin t-testillä erikseen:
  9 *    P1=2*(1-t.F(N-1,t1))                                    P1.=0.3197
 10 *Vastaava F-testi:
 11 *    PF=1-F.F(1,N-1,t1*t1)                                   PF.=0.3197
 12 *T2-testi:
 13 *    T2(r):=1/(1-r*r)*(t1*t1+t2*t2-2*r*t1*t2)
 14 *T2-testin hylkäystodennäköisyys eri korrelaatiokertoimen r arvoilla:
 15 *    P2(r):=1-F.F(p,N-p,(N-p)/(N-1)/p*T2(r))
 16 *                                                  P2(+0.7).=0.5605
 17 *                                                  P2(-0.7).=0.041
 18 *                                                  P2(+0.0).=0.3753
 19 *
```

Otamme toiseksi esimerkiksi simuloidun kolmen muuttujan ja 1000 havainnon otoksen, joka luotiin edellä luvussa *Multinormaalisen otoksen simulointi*. Seuraava Survon laskentakaavio osoittaa, miten T^2-testin avulla tarkastetaan, että otoksesta laskettu keskiarvovektori on sopusoinnussa generoinnin lähtökohtana olleen odotusarvovektorin kanssa.

```
 37  1 SURVO 84C EDITOR Sat Feb 12 14:40:05 1994        D:\M\MONI\ 100 100 0
 30 *
 31 *   T2-testi 3 muuttujan simuloidulle aineistolle
 32 *
 33 *   N=1000 p=3
 34 *
 35 *Kovarianssimatriisin S laskeminen:
 36 *MAT D=MSN.M(*,2)    / Erotetaan hajontojen pystyrivi.
 37 *MAT D!=DV(D)        / Muunnetaan se lävistäjämatriisiksi.
 38 *MAT S=D*CORR.M      / *S~D*R(OTOS) 3*3
 39 *MAT S=S*D           / *S~D*R(OTOS)*D 3*3
 40 *
 41 *Nollahypoteesin mukainen odotusarvovektori (0,1,-2) on matriisin M
 42 *ensimmäinen pystyrivi.
 43 *
 44 *T2-testisuureen laskeminen:
 45 *MAT M0=M(*,1)       / Erotetaan odotusarvovektori
 46 *MAT K=MSN.M(*,1)    / Erotetaan keskiarvojen pystyrivi.
 47 *MAT E!=K-M0         / *E~MSN.M(*,1)-M(*,1) 3*1
 48 *MAT T=INV(S)        / *T~INV(D*R(OTOS)*D) 3*3
 49 *MAT T2=T*E          / *T2~INV(D*R(OTOS)*D)*E 3*1
 50 *MAT E=E'            / *E~E' 1*3
 51 *MAT T2=E*T2         / *T2~E'*INV(D*R(OTOS)*D)*E D1*1
 52 *MAT T2=(N)*T2       / *T2~(N)*E'*INV(D*R(OTOS)*D)*E D1*1
 53 *
 54 *MAT LOAD T2,CUR+1
 55 *MATRIX T2
 56 *(N)*E'*INV(D*R(OTOS)*D)*E
 57 *///          mean
 58 *mean     1.519949
 59 *
 60 *1-F.F(p,N-p,(N-p)/(N-1)/p*1.519949)=0.67846441461669
 61 *
```

Sama tehtävä, joka edellä on yksityiskohtaisesti toteutettu Survon matriisitulkin ja editoriaalisen laskennan avulla, on mahdollista suorittaa automaattisesti tätä varten laadittua sukroa `/MTEST-T2/1` käyttäen. `MTEST` on sukrokokoel-

ma, joka sisältää erilaisia monimuuttujatestejä. Nämä testit soveltavat tarvittavia Survon operaatioita, etenkin matriisiketjuja ja editoriaalista laskentaa ja antavat lopputuloksen tiivistetyssä muodossa.

Niinpä edellinen tehtävä toteutetaan yksinkertaisimmin aktivoimalla sukro /MTEST-T2/1 seuraavasti:

```
  1  1 SURVO 84C EDITOR Sun Feb 13 11:06:28 1994          D:\M\MONI\ 100 100 0
 63 *
 64 */MTEST-T2/1 CORR.M,MSN.M,M
 65 *Hotelling's one-sample T2 test for the mean vector:
 66 *T2=1.51995 p=3 n=1000
 67 *1-F.F(p,n-p,(n-p)/(n-1)/p*T2)=0.67846
 68 *
```

Sukrokomennon parametreina (rivillä 64) ovat otoksesta laskettu korrelaatiomatriisi CORR.M, keskiarvojen, hajontojen ja havaintoluvun muodostama matriisi MSN.M ja nollahypoteesia vastaava odotusarvovektori talletettuna tässä matriisitiedostoksi M.MAT .

Testaustulokset tulevat komentorivin alapuolelle (rivit 65-67) ja ovat samat kuin edellä saadut.

3.4.3 Hotellingin T^2-testi (kahden otoksen vertailu)

Olkoon

$$\mathbf{x}^{(1)}(i), \mathbf{x}^{(2)}(i), ..., \mathbf{x}^{(N_i)}(i)$$

otos jakaumasta $N(\mu^{(i)},\Sigma)$ ja olkoot otokset i=1,2 toisistaan riippumattomat.

Testataan hypoteesia H_0: $\mu^{(1)} = \mu^{(2)}$, kun perusjakaumille yhteinen kovarianssimatriisi Σ on tuntematon. On luonnollista perustaa testi otoskeskiarvovektorien erotukseen $\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}$.

Nämä keskiarvovektorit ovat toisistaan riippumattomia ja

$$\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)} \sim N(\mathbf{0}, (\frac{1}{N_1} + \frac{1}{N_2})\Sigma) .$$

Olkoon

$$\mathbf{S} = \frac{1}{N_1+N_2-2} (\mathbf{A}^{(1)} + \mathbf{A}^{(2)}) ,$$

missä

$$\mathbf{A}^{(i)} = \sum_{\alpha=1}^{N_i} (\mathbf{x}^{(\alpha)}(i) - \bar{\mathbf{x}}^{(i)})(\mathbf{x}^{(\alpha)}(i) - \bar{\mathbf{x}}^{(i)})' \quad , \; i=1,2 .$$

Tällöin

$$(N_1+N_2-2)\mathbf{S} = (\mathbf{A}^{(1)}+\mathbf{A}^{(2)}) \sim W(N_1+N_2-2,\Sigma)$$

ja $\mathbf{S}$ ja $\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}$ ovat toisistaan riippumattomia satunnaissuureita.

T^2-testikriteeri, analogisesti edellisen tapauksen kanssa, saa tällöin muodon

$$T^2 = \frac{N_1N_2}{N_1+N_2} (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})'\mathbf{S}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) .$$

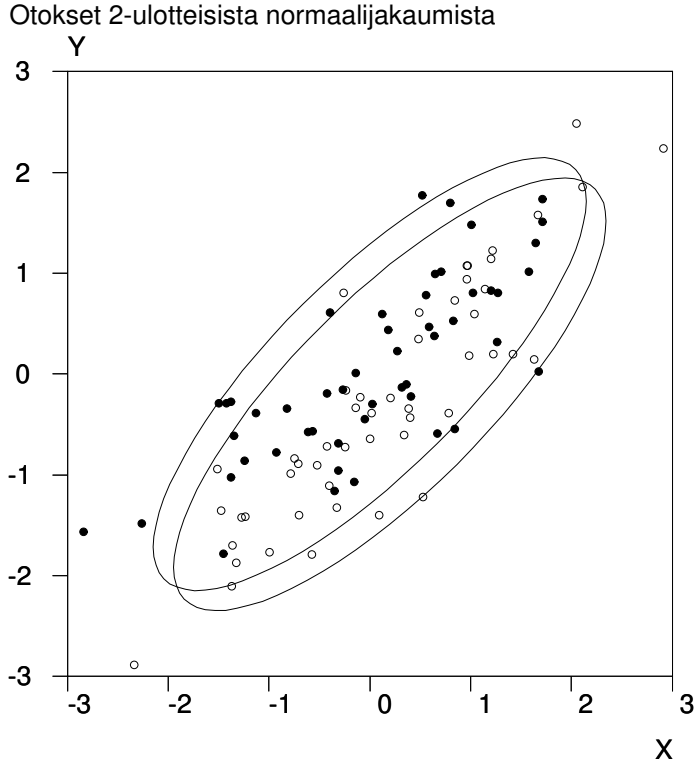
Jos H_0 pätee,

$$\frac{N_1+N_2-p-1}{(N_1+N_2-2)p} T^2 \sim F_{p,N_1+N_2-p-1} .$$

Esimerkkinä tämän testin toimivuudesta luomme kaksi 50 havainnon otosta 2-ulotteisesta normaalijakaumasta. Jakaumilla on sama kovarianssimatriisi mutta eri odotusarvovektorit. Seuraava Survo-kaavio luo nämä otokset:

```
 14  1 SURVO 84C EDITOR Sun Feb 13 17:55:17 1994        D:\M\MONI\ 120 100 0
  1 *
  2 *FILE CREATE N2
  3 *  Kaksi otosta 2-ulotteisista normaalijakaumista::
  4 *  50 ensimmäistä havaintoa: m11=0    m12=0     s11=1 s12=1 r=0.8
  5 *  50 viimeistä   havaintoa: m21=0.2  m22=-0.2  s21=1 s22=1
  6 *FIELDS:
  7 *1 N 4 X
  8 *2 N 4 Y
  9 *END
 10 *
 11 *FILE INIT N2,100
 12 *
 13 *VAR X,Y TO N2
 14 *X=if(ORDER<51)then(X1)else(X2)
 15 *Y=if(ORDER<51)then(Y1)else(Y2)
 16 *X1=Z1     Y1=r*Z1+s*Z2      s=sqrt(1-r*r)
 17 *X2=Z1+0.2 Y2=r*Z1+s*Z2-0.2
 18 *Z1=probit(rand(5))
 19 *Z2=probit(rand(5))
 20 *
```

Kun otokset piirretään XY-koordinaatistoon, voidaan todeta, ettei ero näytä kovin suurelta. Kuvaan on lisätty kummankin perusjakauman osalta hajontaellipsit, joiden sisälle ao. jakauman havainto osuu todennäköisyydellä 0.9.



Kuva on synnytetty seuraavilla `PLOT`- ja `PRINT`-komennoilla:

```
 18  1 SURVO 84C EDITOR Sun Feb 13 18:03:16 1994        D:\M\MONI\ 120 100 0
 21 *.....................................................................
 22 *PLOT N2,X,Y / HEADER=Otokset_2-ulotteisista_normaalijakaumista
 23 *DEVICE=PS,N21.PS   FRAME=1
 24 *SIZE=1000,1000 XDIV=100,800,100 YDIV=100,800,100
 25 *SCALE=-3(1)-1,0:_0,1:_1,2:_2,3:_3
 26 *POINT=0
 27 *IND=ORDER,1,50
 28 *CONTOUR=0.9 BINORM=0,0,1,1,0.8
 29 *.....................................................................
 30 *PLOT N2,X,Y / HEADER=
 31 *DEVICE=PS,N22.PS   FRAME=1
 32 *SIZE=1000,1000 XDIV=100,800,100 YDIV=100,800,100
 33 *SCALE=-3(1)-1,0:_0,1:_1,2:_2,3:_3
 34 *POINT=3
 35 *IND=ORDER,51,100
 36 *CONTOUR=0.9 BINORM=0.2,-0.2,1,1,0.8
 37 *.....................................................................
 38 *PRINT CUR+1,CUR+3
 39 % 1200
 40 - picture N21.PS,*,*
 41 - picture N22.PS,*,*
 42 *
```

Otosten vertailu tapahtuu T^2-testillä laskemalla kummastakin keskiarvot, hajonnat ja korrelaatiokertoimet CORR-operaatiolla, kopioimalla tulosmatriisit ja käyttämällä sukroa /MTEST-T2/2 seuraavasti:

```
 28  1 SURVO 84C EDITOR Sun Feb 13 18:10:59 1994        D:\M\MONI\ 120 100 0
 42 *.....................................................................
 43 *CORR N2,CUR+1  /  IND=ORDER,1,50
 44 *Means, std.devs and correlations of N2  N=50
 45 *Variable  Mean        Std.dev.
 46 *X         0.026603    1.076433
 47 *Y         0.040381    0.893974
 48 *Correlations:
 49 *             X       Y
 50 * X           1.0000  0.7686
 51 * Y           0.7686  1.0000
 52 *
 53 *MAT R1=CORR.M       / *R1~R(N2) S2*2
 54 *MAT MSN1=MSN.M      / *MSN1~MSN(N2) 2*3
 55 *
 56 *CORR N2,CUR+1  /  IND=ORDER,51,100
 57 *Means, std.devs and correlations of N2  N=50
 58 *Variable  Mean        Std.dev.
 59 *X         0.141632    1.103820
 60 *Y        -0.283632    1.182205
 61 *Correlations:
 62 *             X       Y
 63 * X           1.0000  0.8895
 64 * Y           0.8895  1.0000
 65 *
 66 *MAT R2=CORR.M       / *R2~R(N2) S2*2
 67 *MAT MSN2=MSN.M      / *MSN2~MSN(N2) 2*3
 68 *
 69 */MTEST-T2/2 R1,MSN1,R2,MSN2
 70 *Hotelling's two-sample test for equality of mean vectors:
 71 *T2=13.0487 p=2 n1=50 n2=50
 72 *1-F.F(p,n1+n2-p-1,(n1+n2-p-1)/(n1+n2-2)/p*T2)=0.00232
 73 *
```

Hypoteesi odotusarvovektorien samuudesta hylätään näin erittäin selvästi (P=0.00232). Jos keskiarvoja verrataan erikseen tavallisella kaksisuuntaisella t-testillä, joka voidaan tehdä myös em. sukrolla valitsemalla CORR-operaatiossa vain yksi muuttuja kerrallaan, saadaan P-arvoksi X-muuttujalla 0.6 ja Y-muuttujalla 0.13 eli näin tarkasteltuna ero ei paljastu lainkaan.

3.4.4 Kovarianssimatriiseja koskevia testejä

Tarkastelemme aluksi multinormaalijakauman $N(\mu,\Sigma)$ tapauksessa hypoteesin H_0: $\Sigma = \Sigma_0$ testaamista hypoteesia H_1: $\Sigma \neq \Sigma_0$ vastaan, kun μ on tuntematon. Kun otoskoko N on suuri, on käytettävissä testisuure

$$X^2 = Np(\log N - 1) - N\log|\Sigma_0^{-1}\mathbf{A}| + \mathrm{tr}(\Sigma_0^{-1}\mathbf{A}) ,$$

joka noudattaa H_0:n pätiessä asymptoottisesti χ^2-jakaumaa $p(p+1)/2$ vapausasteella. Tämän ja muut tässä jaksossa esitettävät testit on alunperin esittänyt *G.E.P.Box* vuonna 1949 (kts. esim. Anderson 1958 ja Giri 1977).

Survossa tämä testi suoritetaan sukrolla `/MTEST-COV1` seuraavasti. Käytämme jälleen esimerkkiotoksena 3 muuttujan ja 1000 havainnon aineistoa, joka luotiin kohdassa *Multinormaalisen otoksen simulointi*. Havainnollisuuden vuoksi toistamme tässä myös otoksen generoinnin.

```
 28  1 SURVO 84C EDITOR Tue Feb 22 11:06:40 1994        D:\M\MONI\ 100 100 0
  1 *
  2 *MATRIX R
  3 *///  X1   X2   X3
  4 *X1   1    0.4 -0.7
  5 *X2   0.4  1   -0.2
  6 *X3  -0.7 -0.2  1
  7 *
  8 *MATRIX M
  9 *///  Keski  Hajonta
 10 *X1    0       1
 11 *X2    1       2
 12 *X3   -2       0.5
 13 *
 14 *MAT SAVE R
 15 *MAT SAVE M
 16 */MNSIMUL R,M,OTOS,1000  / RND=rand(1)
 17 *
 18 *CORR OTOS,CUR+1
 19 *Means, std.devs and correlations of OTOS  N=1000
 20 *Variable  Mean        Std.dev.
 21 *X1       -0.016924    1.030074
 22 *X2        0.999173    2.011132
 23 *X3       -1.982124    0.484458
 24 *Correlations:
 25 *             X1      X2      X3
 26 * X1           1.0000  0.3699 -0.6876
 27 * X2           0.3699  1.0000 -0.1868
 28 * X3          -0.6876 -0.1868  1.0000
 29 *
 30 */COV S0,R,M
 31 *
 32 */MTEST-COV1 CORR.M,MSN.M,S0_
 33 *H0: Covariance matrix = S0    Test statistics X2=12.211212397085
 34 *1-Chi2.F(df,X2)=0.0574194164358  df=6
 35 *
```

Hypoteettinen (tässä todellinen) Σ_0 (`S0`) on laskettu edellä olevassa kaaviossa sukrolla `/COV` matriiseista **R** ja **M** (rivillä 30). Sukrokomennossa `/MTEST-COV1` (rivillä 32) ovat parametreina otoksesta laskettu korrelaatiomatriisi (`CORR.M`), keskiarvojen ja hajontojen matriisi (`MSN.M`) ja H_0:n mukainen kovarianssimatriisi (`S0`).

Tässä tapauksessa testin *P*-arvo on 0.0574, mikä antaisi aihetta lieviin epäilyksiin H_0:n paikkansapitävyydestä. Tulee kuitenkin muistaa, että yli yksi 20 tällaisesta otoksesta antaa ainakin yhtä "huonon" tuloksen. Kannattaa toistaa

koetta vaihtelemalla satunnaislukugeneraattoria (RND-täsmennys rivillä 16) ja katsoa, miten hyvin testi toimii tässä nollahypoteesin mukaisessa tilanteessa.

On aina mahdollista verrata yksittäisiä korrelaatiokertoimia esim. Fisherin z-muunnokseen perustuvalla testillä. Tässä tapauksessa suurin yksittäinen ero näyttäisi olevan muuttujien X_1 ja X_2 teoreettisen korrelaatiokertoimen (0.4) ja otoksesta saadun (0.3699) välillä. Ero ei kuitenkaan ole merkitsevä, kuten havaitaan seuraavan laskentakaavion avulla.

```
 12  1 SURVO 84C EDITOR Tue Feb 22 12:10:19 1994          D:\M\MONI\ 100 100 0
 37 *
 38 *Fisherin z-muunnos korrelaatiokertoimelle r:
 39 *  z(r):=0.5*log((1+r)/(1-r))
 40 *
 41 *Jos todellinen korrelaatiokerroin on rho0,
 42 *  U=sqrt(n-3)*(z(r)-z(rho0))
 43 *noudattaa likimain normaalijakaumaa N(0,1).
 44 *
 45 *Testattaessa hypoteesia H0: rho=rho0  hypoteesia H1: rho<>rho0 vastaan
 46 *testin kriittinen taso on
 47 *  P=if(r<rho0)then(2*N.F(0,1,U))else(2*(1-N.F(0,1,U)))
 48 *
 49 *Olkoon n=1000 rho0=0.4 ja r=0.3699 .
 50 *Tällöin U.=-1.1159246081983
 51 *        P.=0.26445440729275
 52 *
```

On helppo todeta, että muilla korrelaatiokertoimilla P-arvo on vieläkin suurempi. Se, että koko kovarianssimatriisin vertailussa saadaan merkitsevämpi ero, johtunee tässä tapauksessa siitä, että kaikki korrelaatiokertoimet sattuvat olemaan itseisarvoltaan "oikeita" korrelaatiokertoimia pienempiä, jolloin epäilys nollahypoteesia kohtaan kasautuu.

Verrattaessa useita multinormaalisista otoksista saatuja kovarianssimatriiseja on olemassa vastaavanlainen suurille otoksille tarkoitettu testi (kts. Anderson 1958). Tässä testissä nollahypoteesina on se, että kaikki otokset on saatu multinormaalijakaumista, joissa kovarianssimatriisi Σ on sama, mutta odotusarvovektorien ei tarvitse olla samoja. Survossa tällainen vertailu tehdään sukrolla /MTEST-COV2. Esimerkkinä olemme jakaneet alkuperäisen 1000 havainnon otoksen kolmeen osaan, joissa otoskoot ovat 300, 300 ja 400. Vertaamme näistä osaotoksista laskettuja kovarianssimatriiseja seuraavasti:

```
 12  1 SURVO 84C EDITOR Tue Feb 22 12:39:36 1994          D:\M\MONI\ 100 100 0
 17 *
 18 *CORR OTOS  / IND=ORDER,1,300
 19 */COV COV1
 20 *CORR OTOS  / IND=ORDER,301,600
 21 */COV COV2
 22 *CORR OTOS  / IND=ORDER,601,1000
 23 */COV COV3
 24 *
 25 *COV=COV1,COV2,COV3
 26 *SIZES=300,300,400
 27 */MTEST-COV2
 28 *H0: Equality of covariance matrices    Test statistics X2=8.950502838093
 29 *1-chi2.F(df,X2)=0.70715224310358  df=12
 30 *
```

Sukro /MTEST-COV2 edellyttää, että ao. otoksista on laskettu kovarianssimatriisit (tässä CORR-operaation jälkeen /COV-sukrolla). Vertailtavien kovarianssimatriisitiedostojen nimet annetaan COV-täsmennyksellä (rivi 25) ja otoskoot SIZES-täsmennyksellä (rivi 26). Itse sukrokomennossa ei tarvita mitään parametreja. Tulokset näkyvät riveillä 28-29 ja osoittavat, ettei ole aihetta epäillä nollahypoteesia.

3.4.5 Sama multinormaalijakauma

Edellisen testin yleistyksen avulla voidaan tutkia, ovatko otokset saatu samasta multinormaalijakaumasta. Yleinen hypoteesi on siis, että jokainen otos on multinormaalinen ja nollahypoteesi, että otokset on saatu samasta jakaumasta $N(\mu,\Sigma)$, missä sekä μ että Σ ovat tuntemattomia. Vastahypoteesina on se, että ainakin yksi otoksista on kotoisin jostain muusta multinormaalijakaumasta.
On huomattava, että tarkastelu tapahtuu ehdolla, että kyseessä ovat multinormaaliset otokset; testillä ei siis voi tutkia multinormaalisuutta.

Survossa tämä testi lasketaan sukrolla /MTEST-SAMEMULT, joka on samanrakenteinen kuin edellä käytetty /MTEST-COV2, mutta jossa COV-täsmennyksessä annetaan ensimmäisenä koko yhdistetyn otoksen kovarianssimatriisi ja vastaavasti SIZES-täsmennyksessä tämän yhdistetyn otoksen koko. Äskeinen esimerkki muuntuu tällöin seuraavasti:

```
 16  1 SURVO 84C EDITOR Wed Feb 23 08:25:48 1994        D:\M\MONI\ 100 100 0
 17 *
 18 *CORR OTOS  / IND=ORDER,1,1000
 19 */COV COV
 20 *CORR OTOS  / IND=ORDER,1,300
 21 */COV COV1
 22 *CORR OTOS  / IND=ORDER,301,600
 23 */COV COV2
 24 *CORR OTOS  / IND=ORDER,601,1000
 25 */COV COV3
 26 *
 27 *COV=COV,COV1,COV2,COV3
 28 *SIZES=1000,300,300,400
 29 */MTEST-SAMEMULT_
 30 *H0: Same multivariate normal distr.    Test statistics X2=14.77240369512
 31 *1-chi2.F(df,X2)=0.25411968306762  df=12
 32 *
```

Testin antamat tulokset ovat riveillä 30-31. *P*-arvo 0.254 on huonompi kuin edellisessä testissä, jossa tutkittiin vain kovarianssimatriisien samuutta, mutta nollahypoteesi yhteisestä multinormaalijakaumasta jää selvästi voimaan.

3.4.6 Yksittäisten korrelaatiokertoimien testaaminen

Yksittäisiä multinormaalisesta otoksesta laskettuja korrelaatiokertoimia voi verrata teoreettisiin arvoihin käyttämällä hyväksi Fisherin *z*-muunnosta. Tästä oli esimerkki edellisessä jaksossa.

Jos nollahypoteesina on se, että korrelaatiokerroin ρ on 0, tiedetään, että tämän hypoteesin ollessa voimassa otoksesta lasketulle korrelaatiokertoimelle *r* pätee se, että

$$t = \sqrt{N-2}\ \frac{r}{\sqrt{1-r^2}}$$

noudattaa *t*-jakaumaa *N*-2 vapausasteella.

Kun tutkitaan korrelaatiomatriisia, Survossa on erityinen sukro /LOADCORR korrelaatiomatriisin tulostusta varten siten, että ne korrelaatiokertoimet, jotka em. *t*-testin mielessä ovat merkitsevästi nollasta poikkeavia tulostetaan tehostettuina (värillisinä) vieläpä siten, että riskitasoille *P*=0.001, *P*=0.01 ja *P*=0.05 on kullekin oma tehosteensa.

Tällainen tarkastelu helpottaa mielenkiintoisten korrelaatioiden löytämistä. On kuitenkin syytä huomata, että muuttujamäärän ollessa suuri, merkitsevänttuisia korrelaatiokertoimia ilmestyy luonnostaan sattumalta yllättävän paljon, koska lukuja matriisissa on runsaasti.

Varoittavana esimerkkinä luomme Survolla 100 havainnon otoksen 25 *riippumattomasta* normaalisesta muuttujasta seuraavasti:

```
 25  1 SURVO 84C EDITOR Wed Feb 23 09:09:47 1994        D:\M\MONI\ 120 100 0
  1 *
  2 *p=25
  3 *
  4 *MAT R=IDN(p,p)
  5 *MAT RLABELS X TO R
  6 *MAT CLABELS X TO R
  7 *MAT MS=ZER(p,2)
  8 *MAT Y1=CON(p,1)
  9 *MAT MS(1,2)=Y1
 10 *MAT RLABELS X TO MS
 11 *MAT MS(0,1)="Keski"
 12 *MAT MS(0,2)="Hajonta"
 13 *
 14 */MNSIMUL R,MS,KOE100,100_     / RND=rand(3)
 15 *
```

Otos syntyy /MNSIMUL-sukrolla, jolle annetaan parametreiksi korrelaatiomatriisi R=I muuttujista X_1,X_2,...,X_{25} ja matriisi MS, jonka ensimmäinen sarake "Keski" on keskiarvojen muodostama nollista koostuva vektori ja toinen sarake "Hajonta" koostuu keskihajonnoista =1. Kyseiset matriisit on rakennettu riveillä 4-12 olevilla matriisikäskyillä. /MNSIMUL arpoo tältä pohjalta 100 havainnon otoksen Survon havaintotiedostoksi KOE100.

Koetta jatketaan nyt (seuraava kaavio) laskemalla otoksesta KOE100 keskiarvot, hajonnat ja korrelaatiokertoimet. Viimeksimainitut tallentuvat automaattisesti matriisitiedostoon CORR.M . Koska CORR-komennossa ei ole annettu tulostusriviä, Survon toimituskenttään ei tässä vaiheessa ilmaannu mitään tuloksia.

Rivin 17 REDIM-komennolla tehdään riittävästi tilaa tulostettavalle korrelaatiomatriisille ja erityisesti varataan paikat 80 (värilliselle) varjoriville.

Varsinainen tulostus (riveille 19-103) syntyy yksinkertaisesti pelkällä /LOADCORR-komennolla, joka kirjoittaa korrelaatiomatriisin sopivasti lohkottuna toimituskenttään. Ottamalla huomioon otoskoon *N*=100 se laskee kriittisiä tasoja *P*=0.001, *P*=0.01 ja *P*=0.05 vastaavat korrelaatiokertoimien vähimmäisarvot ja ilmoittaa ne tässä rivillä 21 tehostuksineen. Matriisi tulos-

tetaan tämän jälkeen näiden sopimusten mukaisesti.

```
 10  1 SURVO 84C EDITOR Wed Feb 23 09:11:08 1994        D:\M\MONI\ 120 100 0
 15 *
 16 *CORR KOE100
 17 *REDIM 120,100,80
 18 */LOADCORR
 19 *.....................................................................
 20 *LIMITS=-0.324,-0.253,-0.196,0.196,0.253,0.324,1 SHADOWS=7,1,6,0,6,1,7
 21 *Limits: P=0.001 0.324   P=0.01 0.253   P=0.05 0.196
 22 *LOADM CORR.M,12.123,CUR+1
 23 *R(KOE100)
 24 *              X1     X2     X3     X4     X5     X6     X7     X8     X9
 25 *X1         1.000  0.072 -0.103 -0.051 -0.094 -0.063 -0.110 -0.106  0.069
 26 *X2         0.072  1.000  0.225  0.048  0.003 -0.051  0.000 -0.033  0.160
 27 *X3        -0.103  0.225  1.000  0.063  0.015  0.104  0.252 -0.034  0.031
 28 *X4        -0.051  0.048  0.063  1.000  0.003 -0.144 -0.033 -0.011  0.076
 29 *X5        -0.094  0.003  0.015  0.003  1.000  0.067 -0.007 -0.049 -0.043
 30 *X6        -0.063 -0.051  0.104 -0.144  0.067  1.000  0.172  0.030  0.075
 31 *X7        -0.110  0.000  0.252 -0.033 -0.007  0.172  1.000  0.028 -0.081
 32 *X8        -0.106 -0.033 -0.034 -0.011 -0.049  0.030  0.028  1.000  0.054
 33 *X9         0.069  0.160  0.031  0.076 -0.043  0.075 -0.081  0.054  1.000
 34 *X10        0.095 -0.048 -0.017 -0.044  0.022 -0.148 -0.015 -0.122  0.067
 35 *X11       -0.018  0.009 -0.072  0.258  0.060 -0.016  0.058 -0.010 -0.021
 36 *X12       -0.113 -0.125 -0.108 -0.068 -0.072 -0.077 -0.014 -0.023 -0.092
 37 *X13       -0.131  0.001 -0.103 -0.184  0.108 -0.099 -0.007 -0.048 -0.006
 38 *X14        0.059  0.133  0.234 -0.151  0.133 -0.025  0.187 -0.018 -0.121
 39 *X15       -0.071  0.218  0.113 -0.105 -0.038  0.082  0.114 -0.077  0.014
 40 *X16        0.141 -0.178  0.043  0.000  0.094 -0.015  0.155  0.064  0.041
 41 *X17        0.064 -0.212 -0.049  0.178 -0.121 -0.023  0.023 -0.118 -0.077
 42 *X18        0.047 -0.090  0.028  0.067 -0.067 -0.103  0.022  0.103 -0.012
 43 *X19        0.015 -0.043 -0.177  0.090 -0.032 -0.084  0.022 -0.028 -0.249
 44 *X20       -0.065 -0.013  0.256  0.238  0.247 -0.015 -0.195 -0.010  0.058
 45 *X21        0.077  0.158  0.101  0.040  0.032  0.007 -0.129  0.085  0.175
 46 *X22        0.121  0.001 -0.053 -0.040  0.010  0.011 -0.018  0.098 -0.051
 47 *X23        0.061  0.038  0.084  0.018 -0.128 -0.227  0.079  0.229 -0.081
 48 *X24        0.065 -0.048 -0.055 -0.118  0.022 -0.051  0.073  0.082  0.029
 49 *X25        0.155 -0.038 -0.056 -0.113  0.043  0.034 -0.038 -0.009 -0.113
 50 *
 51 *             X10    X11    X12    X13    X14    X15    X16    X17    X18
 52 *X1         0.095 -0.018 -0.113 -0.131  0.059 -0.071  0.141  0.064  0.047
 53 *X2        -0.048  0.009 -0.125  0.001  0.133  0.218 -0.178 -0.212 -0.090
 54 *X3        -0.017 -0.072 -0.108 -0.103  0.234  0.113  0.043 -0.049  0.028
 55 *X4        -0.044  0.258 -0.068 -0.184 -0.151 -0.105  0.000  0.178  0.067
 56 *X5         0.022  0.060 -0.072  0.108  0.133 -0.038  0.094 -0.121 -0.067
 57 *X6        -0.148 -0.016 -0.077 -0.099 -0.025  0.082 -0.015 -0.023 -0.103
 58 *X7        -0.015  0.058 -0.014 -0.007  0.187  0.114  0.155  0.023  0.022
 59 *X8        -0.122 -0.010 -0.023 -0.048 -0.018 -0.077  0.064 -0.118  0.103
 60 *X9         0.067 -0.021 -0.092 -0.006 -0.121  0.014  0.041 -0.077 -0.012
 61 *X10        1.000  0.172  0.009  0.162  0.115  0.080  0.009 -0.034  0.021
 62 *X11        0.172  1.000  0.144 -0.049 -0.001  0.068 -0.054 -0.238 -0.084
 63 *X12        0.009  0.144  1.000 -0.033 -0.198  0.058 -0.160 -0.009  0.054
 64 *X13        0.162 -0.049 -0.033  1.000  0.005  0.293  0.041  0.139 -0.003
 65 *X14        0.115 -0.001 -0.198  0.005  1.000 -0.006  0.173 -0.050 -0.131
 66 *X15        0.080  0.068  0.058  0.293 -0.006  1.000 -0.063  0.019  0.075
 67 *X16        0.009 -0.054 -0.160  0.041  0.173 -0.063  1.000  0.122  0.063
 68 *X17       -0.034 -0.238 -0.009  0.139 -0.050  0.019  0.122  1.000 -0.031
 69 *X18        0.021 -0.084  0.054 -0.003 -0.131  0.075  0.063 -0.031  1.000
 70 *X19       -0.006  0.145 -0.047 -0.133 -0.029 -0.142  0.235 -0.074 -0.072
 71 *X20        0.015  0.096 -0.057  0.058  0.006  0.065 -0.066 -0.003 -0.072
 72 *X21        0.181 -0.044 -0.040 -0.091  0.010  0.105  0.112 -0.042  0.083
 73 *X22        0.155 -0.109 -0.120 -0.146  0.093  0.011  0.079  0.119 -0.079
 74 *X23        0.187  0.145  0.074  0.164  0.065  0.054  0.063 -0.047  0.148
 75 *X24       -0.061 -0.151 -0.057  0.057  0.182 -0.008  0.129  0.006  0.058
 76 *X25       -0.050 -0.127  0.057  0.046 -0.083 -0.052  0.011  0.126 -0.048
 77 *
```

Jatkuu...

```
 10  1 SURVO 84C EDITOR Wed Feb 23 09:11:08 1994          D:\M\MONI\ 120 100 0
 77 *
 78 *              X19    X20    X21    X22    X23    X24    X25
 79 *X1          0.015 -0.065  0.077  0.121  0.061  0.065  0.155
 80 *X2         -0.043 -0.013  0.158  0.001  0.038 -0.048 -0.038
 81 *X3         -0.177  0.256  0.101 -0.053  0.084 -0.055 -0.056
 82 *X4          0.090  0.238  0.040 -0.040  0.018 -0.118 -0.113
 83 *X5         -0.032  0.247  0.032  0.010 -0.128  0.022  0.043
 84 *X6         -0.084 -0.015  0.007  0.011 -0.227 -0.051  0.034
 85 *X7          0.022 -0.195 -0.129 -0.018  0.079  0.073 -0.038
 86 *X8         -0.028 -0.010  0.085  0.098  0.229  0.082 -0.009
 87 *X9         -0.249  0.058  0.175 -0.051 -0.081  0.029 -0.113
 88 *X10        -0.006  0.015  0.181  0.155  0.187 -0.061 -0.050
 89 *X11         0.145  0.096 -0.044 -0.109  0.145 -0.151 -0.127
 90 *X12        -0.047 -0.057 -0.040 -0.120  0.074 -0.057  0.057
 91 *X13        -0.133  0.058 -0.091 -0.146  0.164  0.057  0.046
 92 *X14        -0.029  0.006  0.010  0.093  0.065  0.182 -0.083
 93 *X15        -0.142  0.065  0.105  0.011  0.054 -0.008 -0.052
 94 *X16         0.235 -0.066  0.112  0.079  0.063  0.129  0.011
 95 *X17        -0.074 -0.003 -0.042  0.119 -0.047  0.006  0.126
 96 *X18        -0.072 -0.072  0.083 -0.079  0.148  0.058 -0.048
 97 *X19         1.000 -0.196 -0.033  0.008 -0.055 -0.010  0.011
 98 *X20        -0.196  1.000  0.125 -0.191  0.007  0.026 -0.198
 99 *X21        -0.033  0.125  1.000  0.017 -0.068  0.111  0.034
100 *X22         0.008 -0.191  0.017  1.000  0.049  0.016  0.032
101 *X23        -0.055  0.007 -0.068  0.049  1.000  0.131 -0.068
102 *X24        -0.010  0.026  0.111  0.016  0.131  1.000  0.024
103 *X25         0.011 -0.198  0.034  0.032 -0.068  0.024  1.000
104 *
```

"Merkitsevät" korrelaatiokertoimet erottuvat värinäytössä huomattavasti paremmin kuin tässä paperille siirretyssä esityksessä. Esim. vihreinä näkyvät laimeimmat korrelaatiot ovat yllä *kursiivilla*.

Korrelaatiomatriisissa on $p(p\text{-}1)/2$=300 erilaista alkiota, kun ykköslävistäjää ei oteta lukuun. Tämä merkitsee sitä, että esim. tasolle P=0.01 yltäviä korrelaatiokertoimia on keskimäärin 3 ja tasolle P=0.05 jopa 15. Tässä tapauksessa nämä lukumäärät ovat 3 ja 14, mikä hyvin vastaa odotuksia. "Merkitseviä" korrelaatiokertoimia siis syntyy, vaikka mitään riippuvuuksia ei perusjakaumassa todellakaan esiinny.

On siten vaarallista lähteä tekemään päätelmiä pelkän korrelaatiomatriisin avulla. Jos esim. tositilanteessa tason P=0.01 ylittäviä kertoimia olisi 9 ja tason P=0.05 ylittäviä 45, niin tiedämme edellisen perusteella, että noin kolmasosa niistä saattaa olla pelkän sattuman aiheuttamia.

Monimuuttujamenetelmien eräänä tehtävänä onkin pyrkiä tiivistämään korrelaatiomatriisiin sisältyvää tietoa siten, että todella merkitsevät piirteet erottuvat satunnaisista riippuvuuksista.

Korrelaatiomatriisi on usein hyvä esittää vektoridiagrammina, kuten näkyy sivuilla 4-5 verkkodokumentissa
http://www.survo.fi/tmp/VectorDiagrams.pdf

4. Pääkomponenttianalyysi

Pääkomponenttianalyysin tehtävä on löytää muuttujien toisistaan riippumattomia lineaarisia yhdistelmiä, jotka keräävät mahdollisimman suuren osan alkuperäisten muuttujien kokonaisvaihtelusta. Teknisenä tavoitteena on usein yksinkertaisesti vähentää tutkittavaa ilmiötä kuvaavien muuttujien lukumäärää mahdollisimman paljon. Pääkomponenttianalyysi voidaan määritellä usealla eri tavalla ja siitä on olemassa erilaisia esitysmuotoja.

4.1 Pääkomponenttien määrääminen I

Etsimme muuttujien $\mathbf{X}=(X_1,X_2,...,X_p)$ sellaista lineaarista yhdistelmää

$$Y = b_1X_1 + b_2X_2 + ... + b_pX_p = \mathbf{b}'\mathbf{X} ,$$

jonka varianssi on mahdollisimman suuri. Tehtävä ei ole mielekäs, ellei kertoimien $b_1,b_2,...,b_p$ absoluuttista suuruutta säädellä jollain tavoin. Teknisesti mukavin rajoitus on asettaa kertoimien neliösumma ykköseksi eli $\mathbf{b}'\mathbf{b}=1$.

Toistaiseksi ei tarvitse olettaa $\mathbf{X}$:n jakaumalta multinormaalisuutta. Olkoon $E(\mathbf{X})=\mu$ ja $\text{cov}(\mathbf{X})=\Sigma\geq 0$. Tehtävänä on siis maksimoida

$$\text{var}(Y)=\mathbf{b}'\text{cov}(\mathbf{X})\mathbf{b}=\mathbf{b}'\Sigma\mathbf{b} \quad \text{ehdolla} \quad \mathbf{b}'\mathbf{b}=1 .$$

Tulos saadaan suoraan matriisin Σ spektraalihajotelmasta,

$$\Sigma = \mathbf{B}\Lambda\mathbf{B}' = \lambda_1\mathbf{b}^{(1)}\mathbf{b}^{(1)'} + \lambda_2\mathbf{b}^{(2)}\mathbf{b}^{(2)'} + ... + \lambda_p\mathbf{b}^{(p)}\mathbf{b}^{(p)'} ,$$

missä Λ on ominaisarvojen $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ lävistäjämatriisi ja $\mathbf{B} = [\mathbf{b}^{(1)},\mathbf{b}^{(2)},...,\mathbf{b}^{(p)}]$ ominaisvektorien $\mathbf{b}^{(1)},\mathbf{b}^{(2)},...,\mathbf{b}^{(p)}$ muodostama ortogonaalinen matriisi.

Matriisihajotelmia koskevassa liitteessä esitetyn tuloksen mukaan neliömuoto $\mathbf{b}'\Sigma\mathbf{b}$ saavuttaa maksimiarvon λ_1, kun $\mathbf{b}=\mathbf{b}^{(1)}$. Muuttujaa $Y_1=\mathbf{b}^{(1)'}\mathbf{X}$ sanotaan 1. pääkomponentiksi ja sen varianssi on siis suurin ominaisarvo λ_1.

Tehtävää voidaan jatkaa etsimällä uutta yhdistettyä muuttujaa $Y=\mathbf{b}'\mathbf{X}$, joka on korreloimaton ensimmäisen pääkomponentin Y_1 kanssa ja jonka varianssi ehdolla $\mathbf{b}'\mathbf{b}=1$ on maksimaalinen. On ilmeistä, että tällöin $Y=Y_2=\mathbf{b}^{(2)'}\mathbf{X}$ ja sen varianssi on λ_2.

Yleisesti *i*. pääkomponentti on $Y=\mathbf{b}'\mathbf{X}=\mathbf{b}^{(i)'}\mathbf{X}=Y_i$ ja sillä on maksimaalinen varianssi ehdoilla $\mathbf{b}'\mathbf{b}=1$ ja

$$\rho(Y,Y_k) = 0 , \quad k=1,2,...,i-1 .$$

Yleensä sovelluksissa tarvitaan vain muutamia ensimmäisiä pääkomponentteja, mutta joskus viimeinenkin eli $Y_p=\mathbf{b}^{(p)'}\mathbf{X}$ voi olla mielenkiintoinen, sillä se ehdolla $\mathbf{b}'\mathbf{b}=1$ *minimoi* yhdistetyn muuttujan $Y=\mathbf{b}'\mathbf{X}$ varianssin.

4.2 Pääkomponenttien määrääminen II

Toinen tapa johtaa pääkomponentit perustuu geometriseen ajatteluun. Tässä yhteydessä multinormaalisuusolettamus on välttämätön. Koska huomio kohdistuu yksinomaan muuttujien väliseen riippuvuuteen, odotusarvot voidaan yksinkertaisuuden vuoksi olettaa nolliksi. Olkoon siis $\mathbf{X} \sim N(\mathbf{0},\Sigma)$, jolloin jakauman tiheysfunktio on

$$n(\mathbf{x} \mid \mathbf{0},\Sigma) = (2\pi)^{-p/2}\det(\Sigma)^{-1/2}\exp(-\tfrac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x}) .$$

p-ulotteiset hajontaellipsoidit $\mathbf{x}'\Sigma^{-1}\mathbf{x}=c$, missä c on positiivinen vakio, määräävät täysin muuttujien $\mathbf{X}$ yhteisvaihtelun luonteen. Pääkomponentit vastaavat näiden ellipsoidien pääakseleita. Tämä todetaan seuraavasti:

Pääakselin päätepisteessä $\mathbf{x}$, vektorin $\mathbf{x}$ pituus $\|\mathbf{x}\|$ saa paikallisen ääriarvon. Nämä potentiaaliset ääriarvopisteet saadaan maksimoimalla $\|\mathbf{x}\|^2=\mathbf{x}'\mathbf{x}$ ehdolla $\mathbf{x}'\Sigma^{-1}\mathbf{x}=c$. Tätä tehtävää vastaa kääntäen neliömuodon $\mathbf{x}'\Sigma\mathbf{x}$ maksimointi ehdolla $\mathbf{x}'\mathbf{x}$=vakio eli pääakselien pituudet $\|\mathbf{x}\|$ ovat verrannollisia kovarianssimatriisin Σ ominaisarvojen $\lambda_1,\lambda_2,...,\lambda_p$ neliöjuuriin.

Sama asia havaitaan tekemällä muunnos $\mathbf{u}=\mathbf{B}'\mathbf{x}$ (kun $\Sigma=\mathbf{B}\Lambda\mathbf{B}'$), joka vastaa koordinaatiston kiertoa. Tämä muunnos saattaa ellipsoidit $\mathbf{x}'\Sigma^{-1}\mathbf{x}=c$ tuttuun pääakselimuotoon $\mathbf{u}'\Lambda^{-1}\mathbf{u}=c$ eli

$$\frac{u_1^2}{\lambda_1} + \frac{u_2^2}{\lambda_2} + ... + \frac{u_p^2}{\lambda_p} = c ,$$

mistä jälleen voidaan päätellä, että i. pääakselin pituuden neliö on verrannollinen ominaisarvoon λ_i, joka on sama kuin i. pääkomponentin Y_i varianssi.

4.2.1 Kahden muuttujan pääakselit ja hajontaellipsit

Aikaisemmin todettiin, että $(\mathbf{X}-\mu)'\Sigma^{-1}(\mathbf{X}-\mu)$ noudattaa χ^2-jakaumaa p vapausasteella. Luottamustasolla P hajontaellipsin yhtälö on siten

$$(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu) = \chi_p^2(P) .$$

Vaativahko harjoitustehtävä on näyttää, että 2 muuttujan tapauksessa tämän ellipsin yhtälö on kirjoitettavissa kätevässä parametrimuodossa

$$x_1 = \mu_1 + \sigma_1 \sqrt{-2\log(1-P)}\,\cos(t) ,$$

$$x_2 = \mu_2 + \sigma_2 \sqrt{-2\log(1-P)}\,\sin(t+\arcsin(\rho)) ,\ 0 \le t \le 2\pi ,$$

josta selvästi näkyy, miten ellipsi riippuu jakauman perusparametreista. Esim. kun $\rho=0$ ja $\sigma_1=\sigma_2$, palaudutaan ympyrän napakoordinaattiesitykseen.

Pääakseleiden yhtälöt ovat kirjoitettavissa esim. muodossa

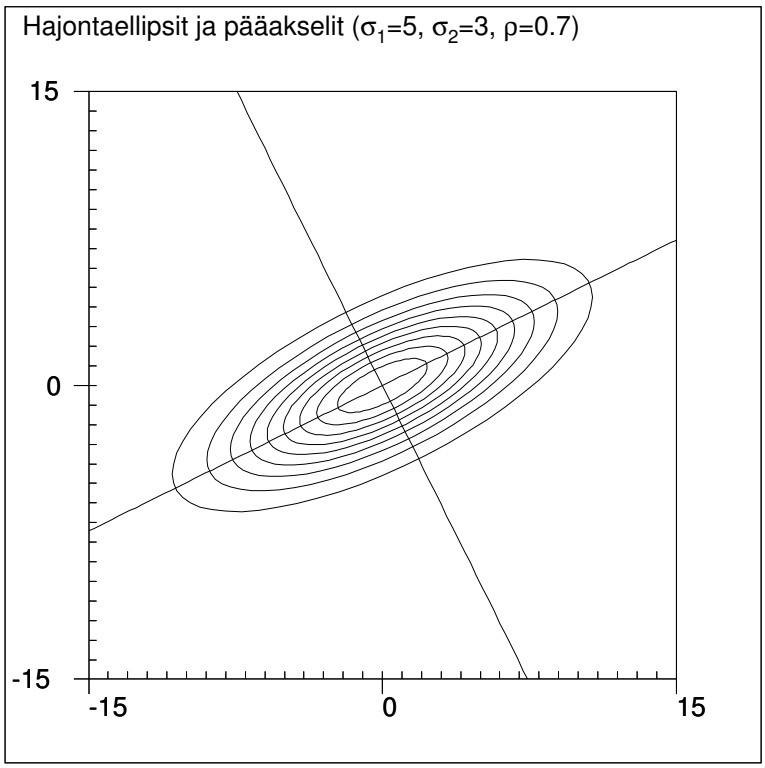
$$x_2 = \mu_2 + \tan(u + k\pi/2)(x_1 - \mu_1), \quad k = 0,1 \ ,$$

missä

$$u = \tfrac{1}{2}\arctan[2\sigma_1\sigma_2\rho/(\sigma_1^2 - \sigma_2^2)] \ .$$

Seuraavat Survon piirroskaaviot piirtävät ao. käyrät kuvaruutuun:

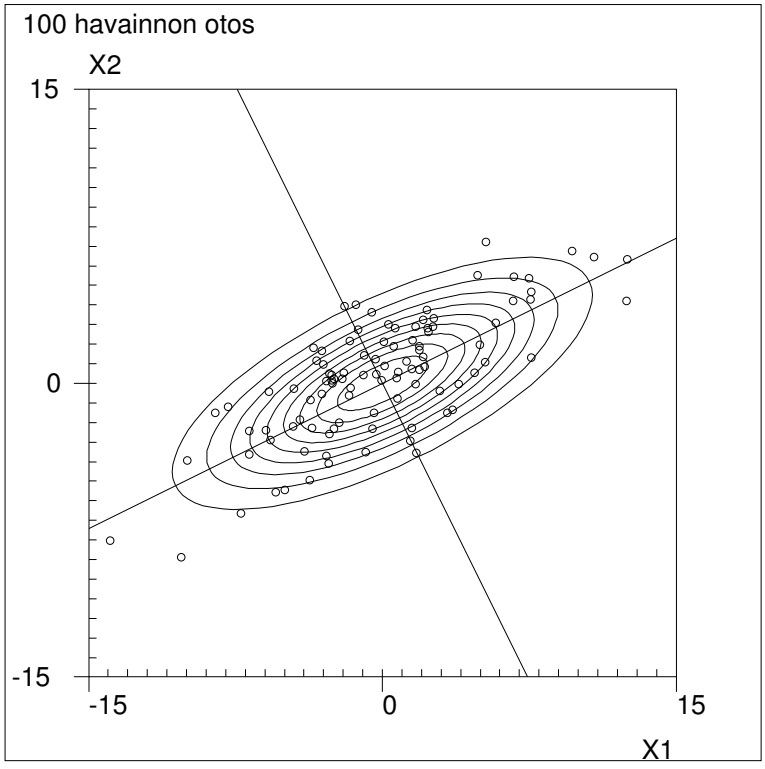
```
 27  1 SURVO 84C EDITOR Mon Feb 28 17:12:22 1994          D:\M\MEN\ 100 100 0
  1 *
  2 *HEADER=Hajontaellipsit_ja_pääakselit_(σ1=5,_σ2=3,_ρ=0.7)
  3 *SIZE=479,479 XDIV=1,8,1 YDIV=1,8,1 SCALE=-15,0,15 TICK=1,1 MODE=VGA
  4 *p=0.1,0.9,0.1                                          *GLOBAL*
  5 *r=0.7 s1=5 s2=3
  6 *GPLOT X(t)=s1*sqrt(-2*log(1-p))*cos(t),
  7 *      Y(t)=s2*sqrt(-2*log(1-p))*sin(t+arcsin(r))
  8 *t=[line_width(2)],0,2*pi,pi/20 pi=3.14159265
  9 *OUTFILE=A
 10 *.....................................................................
 11 *GPLOT Y(X)=X*tan(u+k*pi/2)_
 12 *u=0.5*arctan(2*s1*s2*r/(s1*s1-s2*s2))  k=0,1,1
 13 *INFILE=A
 14 *
```



Vastaavat hajontaellipsoidit ja pääakselit syntyvät tarvittaessa automaattisesti piirrettäessä otosta kaksiulotteisesta jakaumasta. Seuraavassa Survo-esimerkissä on ensin luotu 100 havainnon otos em. jakaumasta ja sitten piirretty se PLOT-kaaviolla, jossa CONTOUR-täsmennys ilmoittaa valitut luottamustasot.

Poikkeuksellisesti arvo 0 tarkoittaa pääakselien piirtämistä. BINORM-täsmennyksellä vuorostaan ilmaistaan jakauman parametrit. Jos se puuttuu, ne estimoidaan otoksesta.

```
 18  1 SURVO 84C EDITOR Mon Feb 28 17:20:51 1994          D:\M\MEN\ 100 100 0
 58 *
 59 *MATRIX R2
 60 *///  X1   X2
 61 *X1   1    0.7
 62 *X2   0.7  1
 63 *
 64 *MATRIX MS2
 65 *///  Keski  Hajonta
 66 *X1     0      5
 67 *X2     0      3
 68 *
 69 *MAT SAVE R2
 70 *MAT SAVE MS2
 71 */MNSIMUL R2,MS2,OTOS2,100    /  RND=rand(1994)
 72 *
 73 *HEADER=100_havainnon_otos
 74 *SIZE=479,479 XDIV=1,8,1 YDIV=1,8,1 SCALE=-15,0,15 TICK=1,1 MODE=VGA
 75 *GPLOT OTOS2,X1,X2_
 76 *CONTOUR=0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9
 77 *BINORM=0,0,5,3,0.7
 78 *
```



Huomattakoon, että uloimman, luottamustasoa 0.9 vastaavan ellipsin ulkopuolelle on joutunut 9 tai 10 havaintoa, mikä vastaa odotettua määrää.

4.3 Pääkomponenttien ominaisuuksia

Edellä esitetyn perusteella muuttujien **X**, joilla E(**X**)=μ ja cov(**X**)=Σ=**B**Λ**B**', pääkomponentit ovat **Y**=**B**'(**X**-μ). Tällöin E(**Y**)=**0** ja cov(**Y**)=Λ, missä lävistäjämatriisin Λ muodostavat Σ:n ominaisarvot $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ ilmaisevat pääkomponenttien $Y_1, Y_2, ..., Y_p$ varianssit.

Spektraalihajotelman

$$\Sigma = \mathbf{B}\Lambda\mathbf{B}' = \lambda_1\mathbf{b}^{(1)}\mathbf{b}^{(1)'} + \lambda_2\mathbf{b}^{(2)}\mathbf{b}^{(2)'} + ... + \lambda_p\mathbf{b}^{(p)}\mathbf{b}^{(p)'}$$

perusteella on ilmeistä, että jäännöskovarianssimatriisi, kun pääkomponentin Y_k vaikutus poistetaan, on

$$\mathrm{cov}(\mathbf{X} \mid Y_k=c) = \Sigma - \lambda_k\mathbf{b}^{(k)}\mathbf{b}^{(k)'} .$$

Näytämme tämän toteen, kun k=1. Muilla k:n arvoilla todistus menee aivan samalla tavalla (merkinnät vain hieman mutkistuvat). Olkoon myös μ=0. Tällöin, koska kääntäen **X**=**BY**,

$$X_i = b_i^{(1)}Y_1 + ... + b_i^{(p)}Y_p , \ i=1,2,...,p$$

ja

$$\mathrm{E}(X_i \mid Y_1=c) = b_i^{(1)}c , \ i=1,2,...,p .$$

Edelleen

$$\mathrm{cov}(X_i,X_j \mid Y_1=c) = \mathrm{E}[(X_i - b_i^{(1)}c)(X_j - b_j^{(1)}c) \mid Y_1=c]$$

$$= \mathrm{E}[(b_i^{(2)}Y_2 + ... + b_i^{(p)}Y_p)\,(b_j^{(2)}Y_2 + ... + b_j^{(p)}Y_p)]$$

$$= \lambda_2 b_i^{(2)}b_j^{(2)} + ... + \lambda_p b_i^{(p)}b_j^{(p)}$$

eli yhteisesti kaikilla i,j-yhdistelmillä kirjoitettuna

$$\mathrm{cov}(\mathbf{X} \mid Y_1=c) = \lambda_2\mathbf{b}^{(2)}\mathbf{b}^{(2)'} + ... + \lambda_p\mathbf{b}^{(p)}\mathbf{b}^{(p)'} = \Sigma - \lambda_1\mathbf{b}^{(1)}\mathbf{b}^{(1)'} .$$

Vastaavasti jos esim. r ensimmäisen pääkomponentin vaikutus eliminoidaan muuttujien X yhteisvaihtelusta, saadaan jäännöskovarianssimatriisiksi

$$\mathrm{cov}(\mathbf{X} \mid Y_i=c_i, \ i=1,2,...,r) = \Sigma - \sum_{i=1}^{r} \lambda_i\mathbf{b}^{(i)}\mathbf{b}^{(i)'} = \sum_{i=r+1}^{p} \lambda_i\mathbf{b}^{(i)}\mathbf{b}^{(i)'}.$$

Edellä kuvattu kokonaisvaihtelun ositus pääkomponenteille oikeuttaa puhumaan "kokonaisvarianssista"

$$\mathrm{tr}(\Sigma) = \lambda_1 + \lambda_2 + ... + \lambda_p$$

ja esim. r ensimmäisen pääkomponentin vaikutuksen poiston jälkeen jäävästä "jäännösvarianssista"

$$\mathrm{tr}[\mathrm{cov}(\mathbf{X} \mid Y_i=c_i, \ i=1,2,...,r)] = \sum_{i=r+1}^{p} \lambda_i \mathrm{tr}[\mathbf{b}^{(i)}\mathbf{b}^{(i)'}] = \lambda_{r+1} + ... + \lambda_p .$$

Täten $\lambda_1 + \lambda_2 + ... + \lambda_r$ on r ensimmäisen pääkomponentin "selittämä varianssi" ja niiden selitysosuus ilmaistaan yleensä prosentteina eli

$$\text{"selitysosuus"} = \frac{\lambda_1 + ... + \lambda_r}{\lambda_1 + ... + \lambda_p} \times 100\% \ .$$

Tavallisesti tyydytään vain osaan pääkomponenteista, jos niiden selitysosuus on riittävä, esim. yli 70%. Usein analyysin lähtökohtana on korrelaatiomatriisi kovarianssimatriisin sijasta, jolloin muuttujien varianssit ikäänkuin samaistetaan ja ainoastaan korrelaatioiden annetaan vaikuttaa pääkomponenttien määräytymiseen. Tällöin kokonaisvaihtelu on sama kuin muuttujien lukumäärä eli $\mathrm{tr}(\Sigma)=p$, jolloin on luonteva pitää niitä pääkomponentteja merkitsevinä, joita vastaavat ominaisarvot ovat (selvästi) yli 1 eli enemmän kuin mikä on kunkin yksittäisen muuttujan varianssi. Erityisen suotavaa on tarkkailla alenevien ominaisarvojen jonoa ja löytää sellainen kohta ykkösen läheisyydestä, jossa tapahtuu selvä putous.

On tietenkin erilaisia ehdotuksia "tilastollisiksi" testeiksi pääkomponenttien lukumäärälle, mutta näillä ei mielestäni ole merkitystä käytännön sovelluksissa. Useimmissa tehtävissä kyseessä on vain muuttujien lukumäärän supistaminen, jolloin "riittävän" selitysosuuden saavuttaminen on paras mitta pääkomponenttien lukumäärälle.

Pääkomponenttianalyysin tulos ilmaistaan tavallisimmin pääkomponenttimatriisina

$$\mathbf{F} = [\ \sqrt{\lambda_1}\ \mathbf{b}^{(1)},\ \ \sqrt{\lambda_2}\ \mathbf{b}^{(2)}, ...,\ \sqrt{\lambda_r}\ \mathbf{b}^{(r)}]\ .$$

Käytämme tästä myös merkintää $P^{(r)}(\Sigma)=\mathbf{F}$. Tässä $P^{(r)}$ tarkoittaa operaatttoria, joka laskee kohteena olevasta matriisista r ensimmäisen pääkomponentin matriisin.

Huomattakoon, että $\mathbf{FF}'=\Sigma$, jos $r=p$ ja $\mathbf{FF}'\approx\Sigma$, jos selitysosuus on "riittävä". Jäännöskovarianssimatriisi voidaan nyt kirjoittaa yksinkertaisesti muodossa $\Sigma - \mathbf{FF}'$.

Pääkomponenttimatriisin normeeraustapaa voidaan perustella seuraavasti. Todetaan, että (kun $\mathrm{E}(\mathbf{X})=\mathbf{0}$)

$$\mathrm{cov}(X_i,Y_j) = \mathrm{E}[(b_i^{(1)}Y_1 + ... + b_i^{(p)}Y_p)Y_j] = b_i^{(j)}\lambda_j$$

eli kun $\mathrm{var}(X_i)=\sigma_i^2$ ja $\mathrm{var}(Y_j)=\lambda_j$, on

$$\rho(X_i,Y_j) = f_{ij}/\sigma_i\ .$$

Erityisesti, kun analyysi tehdään korrelaatiomatriisista eli $\mathbf{F}=P^{(r)}(\mathrm{P})$, on yksinkertaisesti

$$\rho(X_i,Y_j) = f_{ij}$$

eli pääkomponenttimatriisin alkio f_{ij} on suoraan muuttujan X_i ja pääkomponentin Y_j välinen korrelaatiokerroin.

4.4 Pääkomponenttien määrääminen III

Edellä esitetyn perusteella on ymmärrettävää, että myös seuraava matriisiapproksimointiin liittyvä tehtävä on yhdenvertainen pääkomponenttianalyysin kanssa.

Pyritään approksimoimaan kovarianssimatriisi Σ r-asteisella ($r<p$) matriisilla $\mathbf{FF'}$, missä $\mathbf{F}$ on $p\times r$-matriisi. Väitämme, että $\|\Sigma - \mathbf{FF'}\|^2$ saa minimiarvon, kun $\mathbf{F}=P^{(r)}(\Sigma)$. Käytämme tässä Frobeniuksen matriisinormia eli minimoitavana on erotusmatriisin $\Sigma - \mathbf{FF'}$ alkioiden neliösumma.

Todistus perustuu suoraan (liitteen 2) tulokseen, että $\|\mathbf{A} - \mathbf{B}\|^2$ on minimi r-asteisen matriisin $\mathbf{B}$ suhteen, kun $\mathbf{B}=\mathbf{UD}^{(r)}\mathbf{V'}$, missä $\mathbf{UDV'}=\mathbf{A}$ on matriisin $\mathbf{A}$ singulaariarvohajotelma ja $\mathbf{D}^{(r)}$ singulaariarvojen lävistäjämatriisi $\mathbf{D}$ muunnettuna siten, että ensimmäiset r lävistäjäalkiota ovat $d_1, d_2, ..., d_r$, mutta loput asetetaan nolliksi.

Väite seuraa suoraan tästä yleisestä tuloksesta valitsemalla $\mathbf{A}=\Sigma=\mathbf{UDU'}$ ja $\mathbf{F}=\mathbf{U}(\mathbf{D}^{(r)})^{1/2}$. Aikaisempia merkintöjä mukaillen $\mathbf{F}=\mathbf{B}(\Lambda^{(r)})^{1/2}$ eli $\mathbf{F}=P^{(r)}(\Sigma)$.

On syytä huomata, ettei tällä matriisiapproksimoinnilla saatu $\mathbf{F}$ ole yksikäsitteinen, vaan ratkaisuun sisältyy ns. rotaatiomahdollisuus, jonka kohtaamme jatkossa faktorianalyysin yhteydessä. On helppo todeta, että jokainen $\mathbf{F}^*=\mathbf{FT}$, missä $\mathbf{T}$ on ortogonaalinen $r\times r$-matriisi (faktorianalyysin kielellä "ortogonaalinen rotaatiomatriisi") minimoi lausekkeen $\|\Sigma - \mathbf{F}^*\mathbf{F}^{*'}\|^2$, sillä

$$\mathbf{F}^*\mathbf{F}^{*'} = \mathbf{FTT'F'} = \mathbf{FF'} .$$

Yllä valitulle pääkomponenttimatriisille $\mathbf{F}$ on kuitenkin ominaista, että

$$\mathbf{F'F}=(\mathbf{D}^{(r)})^{1/2}\mathbf{U'U}(\mathbf{D}^{(r)})^{1/2} = \mathbf{D}^{(r)} = \Lambda^{(r)}$$

eli matriisi $\mathbf{F}$ on pystyriveittäin ortogonaalinen, mitä "rotatoidut" matriisit $\mathbf{F}^*$ eivät välttämättä ole.

4.5 Pääkomponenttien estimointi ja laskeminen käytännössä

Pääkomponenttien suurimman uskottavuuden estimaatit saadaan analogiaperiaatteella käyttämällä kovarianssimatriisin Σ paikalla sen estimaattoria $\mathbf{A}/N$. Käytännön sovelluksissa on kuitenkin otettava huomioon, että analyysin tulos riippuu ratkaisevasti ja hyvin mutkikkaalla tavalla muuttujien mittayksiköistä. Siten estimoitu kovarianssimatriisi sellaisenaan tulee kysymykseen yleensä vain niissä tilanteissa, joissa *kaikki* muuttujat ovat samanlaatuisia ja mitattu samassa mittayksikössä. Näin on asianlaita yleensä vain joissain luonnontieteellisissä tutkimuksissa. Yhteiskunta- ja käyttäytymistieteissä tarkasteltavaa ilmiötä joudutaan kuvaamaan hyvin erilaatuisilla ja -tasoisilla mittareilla, jolloin mittayksikköjen vertailukelpoisuuden saavuttaminen on lähes mahdotonta. Tavanomainen kompromissi on tällöin siirtyä käyttämään otoksesta estimoitua korrelaatiomatriisia **R** analyysin lähtökohtana. Tämä merkitsee muuttujien varianssien vakiointia, jolloin vain niiden keskinäiset riippuvuudet vaikuttavat tulokseen.

Edelleenkin tutkija voi muuttujavalinnoillaan ratkaisevasti vaikuttaa siihen, mitä analyysista saadaan ulos. Esim. jos tiettyä ominaisuutta mitataan usealla lähisukuisella muuttujalla, voi olla melko varma, että näiden muuttujien kautta tulokseen ilmestyy voimakas pääkomponentti. Samanlainen varaus voidaan tietenkin esittää muidenkin monimuuttujamenetelmien yhteydessä. Kärjistetysti voi jopa sanoa, että varsinkin harkitsemattomilla muuttujavalinnoilla tutkija tutkii enemmän omaa mielikuvaansa ilmiöstä kuin ilmiötä itseään.

Survossa pääkomponenttianalyysin laskelmat voi tehdä suoraan matriisitulkilla. Perussovelluksia varten on kuitenkin saatavilla helppokäyttöiset sukrot `/PCOMPR` ja `/PCOMPCOV`, joista edellinen tekee analyysin korrelaatiomatriisin pohjalta ja jälkimmäinen aidon kovarianssimatriisin pohjalta.

Ennenkuin kuvaamme noiden sukrojen soveltamista, toteamme, että kaikki olennaiset pääkomponenttianalyysiin liittyvät tulokset ovat saatavissa esim. korrelaatiomatriisin tapauksessa suoraan ns. standardoidun havaintomatriisin **Z** singulaariarvohajotelmasta. Käytämme seuraavia aikaisemmin esiteltyjä merkintöjä:

$\mathbf{X}$	alkuperäinen $p \times N$ havaintomatriisi,
$\mathbf{X}$	keskiarvovektorien muodostama $p \times N$-matriisi,
$\mathbf{R}$	korrelaatiomatriisi ($p \times p$),
$\mathbf{D}_s$	keskihajontojen muodostama lävistäjämatriisi ($p \times p$).

Standardoitu havaintomatriisi **Z** on tällöin

$$\mathbf{Z} = \mathbf{D}_s^{-1}(\mathbf{X} - \mathbf{X})/\sqrt{N\text{-}1} \ ,$$

eli korrelaatiomatriisi **R** saadaan lasketuksi suoraan tulona

$$\mathbf{ZZ}' = \mathbf{D}_s^{-1}(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})'\mathbf{D}_s^{-1}/(N\text{-}1) = \mathbf{R}\ .$$

Olkoon nyt $\mathbf{Z}'=\mathbf{UDV}'$ matriisin $\mathbf{Z}'$ singulaariarvohajotelma. Tällöin

$$\mathbf{R} = \mathbf{ZZ}' = \mathbf{VDU}'\mathbf{UDV}' = \mathbf{VD}^2\mathbf{V}'$$

on korrelaatiomatriisin **R** spektraalihajotelma. Tästä voimme suoraan päätellä, että **V** on pääkomponenttien kerroinvektoreiden **B** estimaatti ja $\mathbf{D}^2$ on pääkomponenttien varianssien (ominaisarvojen) Λ estimaatti. Lisäksi pääkomponenttien arvot, joita kutsutaan usein pääkomponenttipistemääriksi, saadaan estimoiduiksi matriisista $\mathbf{UD}=\mathbf{Z}'\mathbf{V}$ eli kaikki olennaiset tulokset on välittömästi luettavissa standardoidun havaintomatriisin singulaariarvohajotelmasta. Jotta pääkomponenttien varianssit vastaisivat tarkalleen ominaisarvoja Λ, pistemäärämatriisi **UD** tulee vielä kertoa luvulla $\sqrt{N\text{-}1}$.

Pääkomponenttianalyysissa ei siis olisi periaatteessa tarvetta lainkaan laskea korrelaatiomatriisia (tai kovarianssimatriisia), koska tulokset pystytään johtamaan suoraan standardoidusta havaintomatriisista singulaariarvohajotelman avulla. Survossa näin tapahtuukin, kun käytetään matriisikomentoa `MATRUN PCOMP` tai Survo-kirjan sivuilla 379-380 kuvattua matriisiketjua. Tämä periaatteessa tarkin ja suorin keino rajoittuu kuitenkin sovelluksiin, joissa koko havaintomatriisi mahtuu kerralla koneen keskusmuistiin.

Tarkastelemme nyt ensiksi pääkomponenttianalyysin suoritusta Survon sukrolla `/PCOMPR`, joka pitää perustanaan aikaisemmin laskettua korrelaatiomatriisia ja näin samaistaa muuttujien varianssit. Käytämme samaa aineistoa ja muuttujavalintaa kuin em. Survo-kirjan esimerkissä. Kymmenottelutiedosto `DECA` on tässä suomennettu tiedostoksi `KYMMEN`.

```
 23  1 SURVO 84C EDITOR Mon Mar 07 11:53:17 1994          D:\M\MEN\ 100 100 0
  1 *
  2 *MASK=--AAAAAAAAA---
  3 *CORR KYMMEN
  4 */PCOMPR CORR.M,MSN.M,4_
  5 *
```

Aluksi lasketaan rivin 3 `CORR`-operaatiolla aineiston `KYMMEN` lajikohtaisista pisteistä korrelaatiot (matriisitiedostoon `CORR.M`) sekä keskiarvot ja hajonnat (tiedostoon `MSN.M`). Muuttujavalinnan osoittaa `MASK`-täsmennys rivillä 2.

Sukro `/PCOMPR` aktivoidaan käyttäen parametreina korrelaatiomatriisia (`CORR.M`), keskiarvojen ja hajontojen matriisia (`MSN.M`) ja haluttua pääkomponenttien lukumäärää (`4`). Jos `/PCOMPR` aktivoidaan ilman parametreja, se kertoo käyttötavastaan. Laskettuaan tulokset uusiin matriisitiedostoihin `/PCOMPR` kirjoittaa toimituskenttään rivit 5-9:

```
  1  1 SURVO 84C EDITOR Mon Mar 07 11:53:30 1994          D:\M\MEN\ 100 100 0
  1 *
  2 *MASK=--AAAAAAAAAA---
  3 *CORR KYMMEN
  4 */PCOMPR CORR.M,MSN.M,4
  5 *MAT LOAD PCOMP.M,END+2 / Principal component loadings
  6 *MAT LOAD PCOMPV.M,END+2 / Variances of principal components
  7 *MAT LOAD PCOCENT.M,END+2 / Variances of components (percentages)
  8 *Use PCOMP.M for factor rotation etc.
  9 *and PCOEFF.M for scores by LINCO <data>,PCOEFF.M(P1,P2,...)
 10 *_
```

Käyttäjälle tarjotaan mahdollisuutta aktivoida mitkä tahansa rivien 5-7 MAT LOAD-komennoista, jotka siirtävät tulosmatriisit näkyville toimituskenttään. Jos nämä kaikki aktivoidaan esitetyssä järjestyksessä, saadaan esiin tulosmatriisit

```
  1  1 SURVO 84C EDITOR Mon Mar 07 11:53:54 1994          D:\M\MEN\ 100 100 0
 11 *MATRIX PCOMP.M
 12 *Principal_components
 13 *///         PCOMP1   PCOMP2   PCOMP3   PCOMP4
 14 *M100       0.22937 -0.81856  0.16769  0.03552
 15 *Pituush   -0.01962 -0.45105 -0.71348 -0.21559
 16 *Kuula     -0.78217 -0.20376  0.26706  0.13231
 17 *Korkeus   -0.49058  0.38403 -0.29577 -0.06899
 18 *M400       0.63734 -0.41117 -0.08840  0.32727
 19 *Aidat     -0.02610 -0.64951 -0.21171  0.07459
 20 *Kiekko    -0.84639 -0.22277  0.15054  0.04996
 21 *Seiväs     0.26796 -0.01509  0.08554 -0.88911
 22 *Keihäs    -0.25012  0.24124 -0.65422  0.15808
 23 *M1500      0.66255  0.49634  0.00452  0.25764
 24 *
 25 *MATRIX PCOMPV.M
 26 *Variances_of_principal_components
 27 *///         PCOMP1   PCOMP2   PCOMP3   PCOMP4   PCOMP5   PCOMP6   PCOMP7
 28 *Variance 2.602056 2.007813 1.206620 1.067052 0.931538 0.594690 0.569080
 29 *
 30 *MATRIX PCOCENT.M
 31 *Variances_of_pr.components_(in_percentages)
 32 *///            1        2        3        4        5        6        7
 33 *Per_cent  26.0206  20.0781  12.0662  10.6705   9.3154   5.9469   5.6908
 34 *Cumulat.  26.0206  46.0987  58.1649  68.8354  78.1508  84.0977  89.7885
 35 *_
```

Näemme, että 4 ensimmäistä pääkomponenttia selittää 68.8% kokonaisvaihtelusta. Esim. matriisi PCOCENT.M saadaan myös loppuosaltaan esille käyttämällä esim. LOADM-komentoa seuraavasti:

```
  1  1 SURVO 84C EDITOR Mon Mar 07 11:55:01 1994          D:\M\MEN\ 100 100 0
 35 *
 36 *LOADM PCOCENT.M,###.####,CUR+1
 37 *Variances_of_pr.components_(in_percentages)
 38 *               1        2        3        4        5        6        7
 39 *Per_cent  26.0206  20.0781  12.0662  10.6705   9.3154   5.9469   5.6908
 40 *Cumulat.  26.0206  46.0987  58.1649  68.8354  78.1508  84.0977  89.7885
 41 *
 42 *               8        9       10
 43 *Per_cent   5.3868   2.4485   2.3761
 44 *Cumulat.  95.1753  97.6239 100.0000
 45 *_
```

Rivillä 9 annetun ohjeen mukaisesti, pääkomponenttien arvot havainnoittain (pääkomponenttipistemäärät) on mahdollista laskea LINCO-operaatiolla käyttämällä /PCOMPR-sukron tulosmatriisia PCOEFF.M. Tähän matriisiin on talletettu sellaiset muuttujien painokertoimet, jotka antavat pääkomponenttien keskiarvoiksi 0 ja variansseiksi korrelaatiomatriisin ominaisarvot alkuperäi-

sessä aineistossa. Näin pääkomponenttien vaihtelun suuruus näkyy oikeudenmukaisesti pääkomponenttipistemäärissä.

LINCO-operaatiossa on ilmoitettu pääkomponentit talletettaviksi muuttujien P1, P2, P3 ja P4 arvoiksi. Ko. muuttujat lisätään havaintotiedostoon KYMMEN automaattisesti, ellei niitä jo ole aikaisemmin perustettu.

Kun pääkomponenttipistemäärät on laskettu LINCO-operaatiolla, tulos on tarkistettu laskemalla sekä alkuperäisten että pääkomponenttimuuttujien keskiarvot, hajonnat ja korrelaatiot CORR-operaatiolla. Rivin 47 MASK-täsmennys ilmoittaa tämän muuttujavalinnan.

Tällöin toimituskenttään tulevat seuraavat keskiarvot ja hajonnat

```
 18  1 SURVO 84C EDITOR Mon Mar 07 12:40:18 1994          D:\M\MEN\ 100 100 0
 45 *.....................................................................
 46 *LINCO KYMMEN,PCOEFF.M(P1,P2,P3,P4)
 47 *MASK=--AAAAAAAAAA---AAAA
 48 *CORR KYMMEN,CUR+1
 49 *Means, std.devs and correlations of KYMMEN  N=48
 50 *Variable  Mean        Std.dev.
 51 *M100      828.1875    59.30256
 52 *Pituush   840.1875    50.72859
 53 *Kuula     740.7708    61.82757
 54 *Korkeus   805.8542    64.80511
 55 *M400      813.5000    49.80216
 56 *Aidat     852.8750    54.20474
 57 *Kiekko    747.4583    62.28212
 58 *Seiväs    900.2708    63.04296
 59 *Keihäs    760.0208    63.93697
 60 *M1500     554.6250    76.67245
 61 *P1       -0.000000    1.613089
 62 *P2       -0.000000    1.416973
 63 *P3       -0.000000    1.098463
 64 *P4        0.000000    1.032982
```

ja korrelaatiokertoimet

```
  1  1 SURVO 84C EDITOR Mon Mar 07 12:53:06 1994          D:\M\MEN\ 100 100 0
 65 *Correlations:
 66 *             M100    Pituush Kuula   Korkeus M400    Aidat   Kiekko
 67 * M100         1.0000  0.1720 -0.0280 -0.4117  0.4561  0.3160  0.0143
 68 * Pituush      0.1720  1.0000 -0.0344 -0.0033  0.1335  0.2981  0.0209
 69 * Kuula       -0.0280 -0.0344  1.0000  0.1625 -0.3037  0.0865  0.7273
 70 * Korkeus     -0.4117 -0.0033  0.1625  1.0000 -0.3388 -0.0390  0.2170
 71 * M400         0.4561  0.1335 -0.3037 -0.3388  1.0000  0.1755 -0.3446
 72 * Aidat        0.3160  0.2981  0.0865 -0.0390  0.1755  1.0000  0.0477
 73 * Kiekko       0.0143  0.0209  0.7273  0.2170 -0.3446  0.0477  1.0000
 74 * Seiväs       0.0547  0.0610 -0.2042 -0.1178  0.0066 -0.0735 -0.1818
 75 * Keihäs      -0.2213  0.1537  0.0231  0.1498 -0.1047 -0.1482  0.1356
 76 * M1500       -0.2917 -0.2067 -0.4462 -0.1461  0.3022 -0.2246 -0.5735
 77 * P1           0.2294 -0.0196 -0.7822 -0.4906  0.6373 -0.0261 -0.8464
 78 * P2          -0.8186 -0.4511 -0.2038  0.3840 -0.4112 -0.6495 -0.2228
 79 * P3           0.1677 -0.7135  0.2671 -0.2958 -0.0884 -0.2117  0.1505
 80 * P4           0.0355 -0.2156  0.1323 -0.0690  0.3273  0.0746  0.0500
```

jatkuu...

```
  1  1 SURVO 84C EDITOR Mon Mar 07 12:53:06 1994          D:\M\MEN\ 100 100 0
 81 *             Seiväs  Keihäs  M1500   P1      P2      P3      P4
 82 * M100         0.0547 -0.2213 -0.2917  0.2294 -0.8186  0.1677  0.0355
 83 * Pituush      0.0610  0.1537 -0.2067 -0.0196 -0.4511 -0.7135 -0.2156
 84 * Kuula       -0.2042  0.0231 -0.4462 -0.7822 -0.2038  0.2671  0.1323
 85 * Korkeus     -0.1178  0.1498 -0.1461 -0.4906  0.3840 -0.2958 -0.0690
 86 * M400         0.0066 -0.1047  0.3022  0.6373 -0.4112 -0.0884  0.3273
 87 * Aidat       -0.0735 -0.1482 -0.2246 -0.0261 -0.6495 -0.2117  0.0746
 88 * Kiekko      -0.1818  0.1356 -0.5735 -0.8464 -0.2228  0.1505  0.0500
 89 * Seiväs       1.0000 -0.1285  0.0125  0.2680 -0.0151  0.0855 -0.8891
 90 * Keihäs      -0.1285  1.0000 -0.0654 -0.2501  0.2412 -0.6542  0.1581
 91 * M1500        0.0125 -0.0654  1.0000  0.6626  0.4963  0.0045  0.2576
 92 * P1           0.2680 -0.2501  0.6626  1.0000 -0.0000 -0.0000 -0.0000
 93 * P2          -0.0151  0.2412  0.4963 -0.0000  1.0000  0.0000 -0.0000
 94 * P3           0.0855 -0.6542  0.0045 -0.0000  0.0000  1.0000  0.0000
 95 * P4          -0.8891  0.1581  0.2576 -0.0000 -0.0000  0.0000  1.0000
 96 *_
```

Tästä tulostuksesta on helppo tarkastaa, että pääkomponenttien keskiarvot ovat nollia ja varianssit ominaisarvojen suuruiset (esim. P1:llä $1.613089^2 = 2.602056$) sekä keskinäiset korrelaatiokertoimet nollia. Nähdään lisäksi, että alkuperäisten muuttujien ja pääkomponenttien väliset korrelaatiokertoimet ovat juuri samat kuin pääkomponenttimatriisin "lataukset" riveillä 14-23.

Koska alkuperäiset muuttujat on mitattu samalla asteikolla (kymmenottelun kansainvälisen pistetaulukon mukaisesti), olisi perusteltua tehdä analyysi suoraan kovarianssimatriisista. Survossa tämä tapahtuu sukrolla /PCOMPCOV ja sen käyttötapa on täsmälleen sama kuin sukron /PCOMPR, jota käytettiin edellä. /PCOMPCOV muodostaa korrelaatiomatriisin ja hajontojen avulla ensin kovarianssimatriisin ja laskee tämän spektraalihajotelman. Pääkomponenttimatriisi normeerataan tässäkin tapauksessa niin, että sen alkiot ovat muuttujien ja pääkomponenttien korrelaatiokertoimia, mikä helpottaa tuloksen tulkintaa.

Vertailun vuoksi on tässä esitetty eräät tärkeimmät tulokset:

```
  1  1 SURVO 84C EDITOR Mon Mar 07 13:26:50 1994          D:\M\MEN\ 100 100 0
  1 *SAVE EX-PCOM2
  2 *MASK=--AAAAAAAAAA---
  3 *CORR KYMMEN
  4 */PCOMPCOV CORR.M,MSN.M,4
  5 *MAT LOAD PCOMP.M,END+2 / Correlations: variables/components
  6 *MAT LOAD PCOMPV.M,END+2 / Variances of principal components
  7 *MAT LOAD PCOCENT.M,END+2 / Variances of components in percentages
  8 *Use PCOEFF.M for scores by LINCO <data>,PCOEFF.M(P1,P2,...)
  9 *
 10 *MATRIX PCOMP.M
 11 *Principal_components
 12 *///        PCOMP1   PCOMP2   PCOMP3   PCOMP4
 13 *M100     -0.02539 -0.86523 -0.14704 -0.00596
 14 *Pituush  -0.12505 -0.29235 -0.28010  0.52635
 15 *Kuula    -0.77693  0.02324  0.04362 -0.35240
 16 *Korkeus  -0.37448  0.59648  0.11325  0.14898
 17 *M400      0.47553 -0.46943 -0.35913 -0.03593
 18 *Aidat    -0.16665 -0.50160 -0.21995  0.02567
 19 *Kiekko   -0.86212  0.01965  0.01665 -0.15963
 20 *Seiväs    0.23018 -0.19530  0.72174  0.52040
 21 *Keihäs   -0.18879  0.40470 -0.55177  0.56799
 22 *M1500     0.82621  0.35012 -0.10504 -0.23885
 23 *_
```

```
  1  1 SURVO 84C EDITOR Mon Mar 07 13:28:21 1994          D:\M\MEN\ 100 100 0
 23 *
 24 *MATRIX PCOMPV.M
 25 *Variances_of_principal_components
 26 *///        PCOMP1   PCOMP2   PCOMP3   PCOMP4   PCOMP5   PCOMP6   PCOMP7
 27 *Variance 10833.69  7178.04  4181.87  4115.51  3442.95  2290.16  1895.31
 28 *
 29 *MATRIX PCOCENT.M
 30 *Variances_of_principal_components_(in_percentages)
 31 *///            1        2        3        4        5        6        7
 32 *Per_cent  29.0051  19.2178  11.1961  11.0185   9.2178   6.1314   5.0743
 33 *Cumulat.  29.0051  48.2228  59.4190  70.4375  79.6553  85.7867  90.8610
 34 *
```

```
  1  1 SURVO 84C EDITOR Mon Mar 07 13:29:00 1994          D:\M\MEN\ 100 100 0
 45 *.....................................................................
 46 *LINCO KYMMEN,PCOEFF.M(P1,P2,P3,P4)
 47 *MASK=--AAAAAAAAAA---AAAA
 48 *CORR KYMMEN,CUR+1
 49 *Means, std.devs and correlations of KYMMEN  N=48
 50 *Variable  Mean        Std.dev.
 51 *M100      828.1875    59.30256
 52 *Pituush   840.1875    50.72859
 53 *Kuula     740.7708    61.82757
 54 *Korkeus   805.8542    64.80511
 55 *M400      813.5000    49.80216
 56 *Aidat     852.8750    54.20474
 57 *Kiekko    747.4583    62.28212
 58 *Seiväs    900.2708    63.04296
 59 *Keihäs    760.0208    63.93697
 60 *M1500     554.6250    76.67245
 61 *P1       -0.000001    104.0850
 62 *P2       -0.000000    84.72329
 63 *P3        0.000000    64.66738
 64 *P4        0.000000    64.15227
```

Lukuunottamatta 1. pääkomponenttia, tuloksissa on eroja, jotka johtuvat siitä, etteivät muuttujien hajonnat ole samoja. Koska esim. muuttujan M1500 hajonta on jonkin verran suurempi kuin muiden, se saa painokkaamman osuuden jälkimmäisen analyysin 1. pääkomponentissa.

4.5.1 Simulointikoe

Pääkomponentteihin viitattiin jo multinormaalijakauman määritelmässä (3) kohdassa 2.2. Todettiin, että multinormaalinen satunnaisvektori **X** voidaan yksinkertaisimmin konstruoida muodossa

$$\mathbf{X} = \mathbf{SDW} + \mu ,$$

missä $\mathbf{W} \sim N(\mathbf{0},\mathbf{I})$, **D** on lävistäjämatriisi, ja **S** ortogonaalinen matriisi. Tällöin muuttujat **DW** ovat muuttujien **X** pääkomponentteja ja niiden varianssit ovat samat kuin matriisin $\mathbf{D}^2$ lävistäjäalkiot.

Tarkastamme nämä ominaisuudet tutkimalla 8 muuttujan tapausta, joka on vajaa-asteinen siten, että pääkomponentteja on vain 3 eli 5 viimeistä **D**:n lävistäjäalkiota ovat nollia. Tällöin voimme tyytyä 8×3-matriisiin **S**, joka on pystyriveittäin ortogonaalinen. Tällainen **S** voidaan muodostaa "mielivaltaisesta", täysiasteisesta 8×3-matriisista **C** Gram-Schmidt-ortogonalisoinnilla seuraavasti:

```
 43  1 SURVO 84C EDITOR Tue Mar 08 09:32:24 1994          D:\M\MEN\ 200 100 0
  1 *
  2 *MATRIX C
  3 *///  W1   W2   W3
  4 *X1   0.8  0.5  0.5
  5 *X2  -0.8  0.2 -0.2
  6 *X3   0.6  0.6  0.3
  7 *X4   0.5  0.0  0.1
  8 *X5   0.0  0.2  0.0
  9 *X6   0.4  0.2 -0.7
 10 *X7  -0.1 -0.3  0.2
 11 *X8  -0.5  0.9 -0.4
 12 *
 13 *MAT SAVE C
 14 *MAT GRAM-SCHMIDT DECOMPOSITION OF C TO S,B
 15 *
```

Matriisi **C** on siis hajotettu samanmuotoisen pystyriveittäin ortogonaalisen matriisin **S** ja yläkolmiomatriisin **B** tuloksi. Jatkossa käytämme vain matriisia **S**, jota nyt vastaa matriisitiedosto `S` (`S.MAT`).

Valitsemme nyt 3 pääkomponentin variansseiksi luvut 5, 2 ja 1. Annamme ne vaakavektorina `D2`. Näiden lukujen neliöjuurista tehty matriisi vastaa määritelmän (3) mukaista **D**-matriisia. Muodostamme tulomatriisin **SD**, jolloin jakauman kovarianssimatriiksi tulee $\Sigma=(\mathbf{SD})(\mathbf{SD})'$, mitä tässä laskelmassa vastaa matriisi `SIGMA`.

```
 19  1 SURVO 84C EDITOR Tue Mar 08 09:35:47 1994          D:\M\MEN\ 200 100 0
 15 *
 16 *MATRIX D2
 17 *///  W1   W2   W3
 18 *D2    5    2    1
 19 *
 20 *MAT SAVE D2
 21 *MAT TRANSFORM D2 BY sqrt(X#)
 22 *MAT D=DV(D2)       / *D~DV(T(D2_by_sqrt(X#))) D3*3
 23 *MAT SD!=S*D          / *SD~GS(C)*DV(T(D2_by_sqrt(X#))) 8*3
 24 *
 25 *MAT SIGMA!=MMT(SD)_  / *SIGMA~SD*SD' S8*8
 26 *
```

Tarvitsemme vielä odotusarvojen ja hajontojen muodostaman 8×2-matriisin (MS), joka synnytetään seuraavilla matriisikäskyillä. Hajonnat saadaan kovarianssimatriisin SIGMA lävistäjäalkioiden neliöjuurina. Odotusarvot asetamme yksinkertaisesti nolliksi, koska niillä ei ole mitään merkitystä näissä tarkasteluissa.

```
  1  1 SURVO 84C EDITOR Tue Mar 08 09:41:10 1994          D:\M\MEN\ 200 100 0
 26 *
 27 *MAT MS=ZER(8,2)
 28 *MAT H!=VD(SIGMA)    / *H~VD(SIGMA) 8*1
 29 *MAT TRANSFORM H BY sqrt(X#)
 30 *MAT MS(1,2)=H
 31 *MAT MS(0,1)="Keski"
 32 *MAT MS(0,2)="Hajonta"
 33 *
 34 *MAT LOAD MS,CUR+1
 35 *MATRIX MS
 36 *CON&T(H_by_sqrt(X#))
 37 *///         Keski  Hajonta
 38 *X1       0.000000 1.311968
 39 *X2       0.000000 1.224255
 40 *X3       0.000000 1.087592
 41 *X4       0.000000 0.741024
 42 *X5       0.000000 0.226104
 43 *X6       0.000000 1.064743
 44 *X7       0.000000 0.408486
 45 *X8       0.000000 1.302422
 46 *_
```

Kovarianssimatriisista muodostamme lopuksi korrelaatiomatriisin R, jolloin **X**-muuttujien jakauma on täydellisesti kuvattuna matriisien R ja MS avulla.

```
 25  1 SURVO 84C EDITOR Tue Mar 08 09:47:01 1994          D:\M\MEN\ 200 100 0
 46 *
 47 *MAT TRANSFORM H BY 1/X#
 48 *MAT H!=DV(H)        / *H~DV(T(H_by_1/X#)) D8*8
 49 *MAT R=SIGMA*H      / *R~SIGMA*H 8*8
 50 *MAT R!=H*R         / *R~H*SIGMA*H 8*8
 51 *
 52 *MAT LOAD R,##.####,CUR+1_
 53 *MATRIX R
 54 *///           X1      X2      X3      X4      X5      X6      X7      X8
 55 *X1        1.0000 -0.7497  0.9746  0.8377  0.3858  0.3315 -0.4649 -0.2434
 56 *X2       -0.7497  1.0000 -0.6201 -0.9833  0.2730 -0.5495  0.1742  0.7539
 57 *X3        0.9746 -0.6201  1.0000  0.7417  0.5717  0.3706 -0.6234 -0.0269
 58 *X4        0.8377 -0.9833  0.7417  1.0000 -0.0968  0.6049 -0.3330 -0.6230
 59 *X5        0.3858  0.2730  0.5717 -0.0968  1.0000  0.0385 -0.7070  0.7985
 60 *X6        0.3315 -0.5495  0.3706  0.6049  0.0385  1.0000 -0.7330 -0.1081
 61 *X7       -0.4649  0.1742 -0.6234 -0.3330 -0.7070 -0.7330  1.0000 -0.4878
 62 *X8       -0.2434  0.7539 -0.0269 -0.6230  0.7985 -0.1081 -0.4878  1.0000
 63 *
```

Teemme nyt pääkomponenttianalyysin laskelmat sukron /PCOMPCOV avulla käyttäen todellisia jakauman parametreja R ja MS. Näemme, että pääkompo-

nenttien varianssit ovat täsmälleen odotetut. Pääkomponenttimatriisi on tässä tapauksessa skaalattu siten, että alkioina ovat muuttujien ja pääkomponenttien väliset korrelaatiokertoimet. Tämä matriisi on sama kuin aikaisempi kuvausmatriisi SD, jonka vaakarivit on jaettu muuttujien hajonnoilla.

```
  1  1 SURVO 84C EDITOR Tue Mar 08 09:54:23 1994          D:\M\MEN\ 200 100 0
 64 *.....................................................................
 65 */PCOMPCOV R,MS,3
 66 *MAT LOAD PCOMP.M,END+2 / Correlations: variables/components
 67 *MAT LOAD PCOMPV.M,END+2 / Variances of principal components
 68 *MAT LOAD PCOCENT.M,END+2 / Variances of components in percentages
 69 *Use PCOEFF.M for scores by LINCO <data>,PCOEFF.M(P1,P2,...)
 70 *
 71 *MATRIX PCOMP.M
 72 *Principal_components
 73 *///         PCOMP1   PCOMP2   PCOMP3
 74 *X1       -0.89711 -0.34928 -0.27055
 75 *X2        0.96138 -0.26482 -0.07491
 76 *X3       -0.81164 -0.54725 -0.20435
 77 *X4       -0.99270  0.08489  0.08571
 78 *X5        0.00000 -0.98873 -0.14972
 79 *X6       -0.55271 -0.16270  0.81734
 80 *X7        0.36017  0.79012 -0.49598
 81 *X8        0.56480 -0.82070  0.08626
 82 *
 83 *MATRIX PCOMPV.M
 84 *Variances_of_principal_components
 85 *///         PCOMP1   PCOMP2   PCOMP3   PCOMP4   PCOMP5   PCOMP6   PCOMP7
 86 *Variance  5.00000  2.00000  1.00000  0.00000  0.00000 -0.00000 -0.00000
 87 *
 88 *MATRIX PCOCENT.M
 89 *Variances_of_principal_components_(in_percentages)
 90 *///             1        2        3        4        5        6        7
 91 *Per_cent   62.500   25.000   12.500    0.000    0.000   -0.000   -0.000
 92 *Cumulat.   62.500   87.500  100.000  100.000  100.000  100.000  100.000
 93 *
```

Katsomme sitten, mitä tapahtuu, kun luomme 1000 havainnon otoksen tästä multinormaalijakaumasta. Tämä tapahtuu sukrolla /MNSIMUL (rivi 95) ja tarpeelliset tunnusluvut lasketaan CORR-komennolla matriisitiedostoihin CORR.M ja MSN.M . Soveltamalla /PCOMPCOV-sukroa näillä otossuureilla saamme seuraavat tulokset:

```
  1  1 SURVO 84C EDITOR Tue Mar 08 10:20:18 1994          D:\M\MEN\ 200 100 0
 94 *.....................................................................
 95 */MNSIMUL R,MS,PDATA,1000  / RND=rand(28101937)
 96 *CORR PDATA
 97 */PCOMPCOV CORR.M,MSN.M,3
 98 *MAT LOAD PCOMP.M,END+2 / Correlations: variables/components
 99 *MAT LOAD PCOMPV.M,END+2 / Variances of principal components
100 *MAT LOAD PCOCENT.M,END+2 / Variances of components in percentages
101 *Use PCOEFF.M for scores by LINCO <data>,PCOEFF.M(P1,P2,...)
102 *
103 *MATRIX PCOMP.M
104 *Principal_components
105 *///         PCOMP1   PCOMP2   PCOMP3
106 *X1        -0.87412 -0.38891 -0.29097
107 *X2         0.97189 -0.22144 -0.07995
108 *X3        -0.76945 -0.59667 -0.22789
109 *X4        -0.99538  0.03952  0.08749
110 *X5         0.10235 -0.97967 -0.17257
111 *X6        -0.52951 -0.22606  0.81763
112 *X7         0.27375  0.83366 -0.47965
113 *X8         0.62138 -0.78072  0.06612
114 *
115 *MATRIX PCOMPV.M
116 *Variances_of_principal_components
117 *///         PCOMP1   PCOMP2   PCOMP3   PCOMP4   PCOMP5   PCOMP6   PCOMP7
118 *Variance  4.90501  2.04985  1.02018  0.00000  0.00000  0.00000  0.00000
119 *
120 *MATRIX PCOCENT.M
121 *Variances_of_principal_components_(in_percentages)
122 *///             1        2        3        4        5        6        7
123 *Per_cent   61.504   25.703   12.792    0.000    0.000    0.000    0.000
124 *Cumulat.   61.504   87.208  100.000  100.000  100.000  100.000  100.000
125 *
```

Havaitsemme, että sekä ominaisarvot (pääkomponenttien varianssit) että pääkomponenttimatriisi vastaavat hyvin teoreettista lähtökohtaa. Ominaisarvot neljännestä eteenpäin ovat edelleen tarkkaan nollia. Tämä johtuu siitä, että jakauman vajaa-asteisuus säilyy täydellisesti myös simuloinnissa.

Saadaksemme enemmän realistisuutta esimerkkiaineistoomme, lisäämme kuhunkin muuttujaan normaalisen satunnaisvirheen, jonka hajonta vaihtelee muuttujittain arvosta 0.1 arvoon 0.8:

```
 37  1 SURVO 84C EDITOR Tue Mar 08 11:05:09 1994          D:\M\MEN\ 200 100 0
126 *.....................................................................
127 *U1=probit(rand(1994))
128 *U2=probit(rand(1994))
129 *U3=probit(rand(1994))
130 *U4=probit(rand(1994))
131 *U5=probit(rand(1994))
132 *U6=probit(rand(1994))
133 *U7=probit(rand(1994))
134 *U8=probit(rand(1994))
135 *
136 *VAR X1,X2,X3,X4,X5,X6,X7,X8 TO PDATA
137 *X1=X1+0.1*U1
138 *X2=X2+0.2*U2
139 *X3=X3+0.3*U3
140 *X4=X4+0.4*U4
141 *X5=X5+0.5*U5
142 *X6=X6+0.6*U6
143 *X7=X7+0.7*U7
144 *X8=X8+0.8*U8
145 *
```

Kun nyt laskemme muunnetusta otoksesta tunnusluvut ja toistamme pääkomponenttianalyysin, saamme tulokset:

```
  1  1 SURVO 84C EDITOR Tue Mar 08 11:09:30 1994          D:\M\MEN\ 200 100 0
145 *
146 *CORR PDATA
147 */PCOMPCOV CORR.M,MSN.M,3
148 *MAT LOAD PCOMP.M,END+2 / Correlations: variables/components
149 *MAT LOAD PCOMPV.M,END+2 / Variances of principal components
150 *MAT LOAD PCOCENT.M,END+2 / Variances of components in percentages
151 *Use PCOEFF.M for scores by LINCO <data>,PCOEFF.M(P1,P2,...)
152 *
153 *MATRIX PCOMP.M
154 *Principal_components
155 *///         PCOMP1   PCOMP2   PCOMP3
156 *X1        -0.83629 -0.39319  0.35593
157 *X2         0.95733 -0.11221  0.05459
158 *X3        -0.71110 -0.58999  0.29937
159 *X4        -0.88258 -0.04127 -0.04362
160 *X5         0.06735 -0.39318  0.15739
161 *X6        -0.49497 -0.33613 -0.78548
162 *X7         0.15345  0.53038  0.20998
163 *X8         0.62711 -0.76076  0.00732
164 *
165 *MATRIX PCOMPV.M
166 *Variances_of_principal_components
167 *///         PCOMP1   PCOMP2   PCOMP3   PCOMP4   PCOMP5   PCOMP6   PCOMP7
168 *Variance 5.119680 2.547909 1.301848 0.493218 0.260157 0.225162 0.125082
169 *
170 *MATRIX PCOCENT.M
171 *Variances_of_principal_components_(in_percentages)
172 *///             1        2        3        4        5        6        7
173 *Per_cent  50.5797  25.1720  12.8616   4.8727   2.5702   2.2245   1.2357
174 *Cumulat.  50.5797  75.7517  88.6133  93.4861  96.0563  98.2807  99.5165
175 *
```

Näemme, että tämä häirintä ei horjuta uskoa 3 pääkomponenttiin, vaikka selitysosuus on pudonnut täydestä 100 prosentista 88.6 prosenttiin. Varianssi romahtaa arvosta 1.30 arvoon 0.49 siirryttäessä kolmannesta neljänteen pääkomponenttiin. Pääkomponenttimatriisi on edelleen hyvin samanlainen; vain kolmannen komponentin suunta on kääntynyt laskentaprosessissa päinvastaiseksi, mikä on täysin sallittua.

Kokeen viime vaiheessa lisätyt häiriömuuttujat vastaavat faktorianalyysimallissa esiin tulevia muuttujien ominaisvaihtelukomponentteja, jotka ovat toisistaan riippumattomia. Faktorianalyysi pystyy selviytymään niiden haitallisista vaikutuksista paremmin kuin pääkomponenttianalyysi.

5. Faktorianalyysi

Faktorianalyysi muistuttaa monessa suhteessa pääkomponenttianalyysia. Pääasiallisia eroja on kaksi:
1. Muuttujien kokonaisvaihtelu jaetaan kahteen osaan, yhteisvaihteluun ja ominaisvaihteluun.
2. Rotaation (eli tässä yhteydessä ortogonaalisen tai lähes ortogonaalisen lineaarisen muunnoksen) avulla pyritään faktoreissa, jotka vastaavat pääkomponentteja, ns. yksinkertaiseen rakenteeseen (simple structure). Tälle rakenteelle pyritään antamaan tutkittavan ilmiön teoriaan liittyvä tulkinta.

On tapana myös korostaa, että faktorianalyysi on kovarianssipainotteinen, kun taas pääkomponenttianalyysi (käyttäessään alkuperäistä kovarianssimatriisia lähtökohtanaan) on varianssipainotteinen menetelmä.

Faktorianalyysissa ei siis enää ole kyse kokonaisvaihtelun maksimaalisesta siirtämisestä uusille muuttujille, vaan vähäulotteisen piilorakenteen löytämisestä muuttujien korrelaatioiden avulla.

5.1 Faktorianalyysimalli

Yleisesti faktorianalyysissa havaitut muuttujat $X_1,X_2,...,X_p$ esitetään mallin

$$X_i = \mu_i + a_{i1}F_1 + a_{i2}F_2 + ... + a_{ir}F_r + U_i\,,\ \ i = 1,2,...,p$$

eli

$$\mathbf{X} = \mu + \mathbf{AF} + \mathbf{U}$$

avulla. Tässä $p\times r$-matriisi $\mathbf{A}$ on faktorimatriisi. Muuttujat $\mathbf{F}=(F_1,F_2,...,F_r)$ ovat (yhteis)faktoreita ja niitä on (yleensä olennaisesti) vähemmän kuin alkuperäisiä muuttujia eli $r<p$. Muuttujat $\mathbf{U}=(U_1,U_2,...,U_p)$ ovat ominaisfaktoreita.

Malliin liittyvistä suureista on tapana tehdä olettamukset:

1) $\mathbf{X} \sim N(\mu,\Sigma),\ \Sigma>0$,

2) $\mathbf{F} \sim N(\mathbf{0},\Phi)$,

3) $\mathbf{U} \sim N(\mathbf{0},\Psi^2)$, missä Ψ^2 on alkioiden $\psi_1^2, \psi_2^2, ..., \psi_p^2$ lävistäjämatriisi,

4) Ominaisfaktorit $\mathbf{U}$ ovat riippumattomia yhteisfaktoreista $\mathbf{F}$,

5) $r(\mathbf{A})=r$.

Asetetut olettamukset antavat kovarianssimatriisille Σ tietyn rakenteen, joka saadaan esille laskemalla Σ em. suureiden avulla:

$$\Sigma = E[(\mathbf{X}-\mu)(\mathbf{X}-\mu)'] = E[(\mathbf{AF}+\mathbf{U})(\mathbf{AF}+\mathbf{U})'] = E(\mathbf{AFF'A'}) + E(\mathbf{UU}')$$

eli

$$\mathbf{\Sigma} = \mathbf{A\Phi A}' + \Psi^2 .$$

Tätä sanotaan faktorianalyysin perusyhtälöksi.

Faktorit **F** eivät määräydy em. olettamuksista ja perusyhtälöstä yksikäsitteisesti, vaan jäljelle jää ns. rotaatiomahdollisuus. Olkoon $\mathbf{F}^*=\mathbf{T}'\mathbf{F}$, missä **T** on säännöllinen $r\times r$-matriisi, jolloin $\mathbf{F}^* \sim N(\mathbf{0},\mathbf{T}'\Phi\mathbf{T})$. Jos merkitään $\Phi^*=\mathbf{T}'\Phi\mathbf{T}$ ja $\mathbf{A}^*=\mathbf{A}(\mathbf{T}^{-1})'$, saadaan muuttujille **X** esitys

$$\mathbf{X} = \mu + \mathbf{A}^*\mathbf{F}^* + \mathbf{U}$$

ja perusyhtälö tulee muotoon

$$\mathbf{\Sigma} = \mathbf{A\Phi A}' + \Psi^2 = \mathbf{A}^*\Phi^*\mathbf{A}^{*'} + \Psi^2$$

eli jokainen $\mathbf{F}^*=\mathbf{T}'\mathbf{F}$ ja sitä vastaava $\mathbf{A}^*=\mathbf{A}(\mathbf{T}^{-1})'$ on myös mahdollinen ratkaisu.

Syntyy siis rotaatio-ongelma eli teknisesti yhtä hyvien ratkaisujen joukosta on lupa valita sellainen, joka on esim. helpoin tulkita. Kaiken kaikkiaan tehtävänä on määrätä suureet r, Ψ^2, **A** ja **T** . Usein halutaan lisäksi estimoida havainnoittain faktorien arvot eli laskea ns. faktoripistemäärät (factor scores). Tavallisesti tehdään vielä yksinkertaistava olettamus

6) $\Phi = \mathbf{I}$

eli faktorit oletetaan keskenään korreloimattomiksi. Tällöin perusyhtälö tulee muotoon

$$\mathbf{\Sigma} = \mathbf{AA}' + \Psi^2$$

ja rotaatiomatriiseiksi **T** kelpaavat vain ortogonaaliset. Tällöin matriisi **A** antaa muuttujien ja faktorien väliset kovarianssit, sillä

$$\mathrm{cov}(\mathbf{X},\mathbf{F}) = \mathrm{E}[(\mathbf{X}-\mu)\mathbf{F}'] = \mathrm{E}[(\mathbf{AF}+\mathbf{U})\mathbf{F}'] = \mathbf{A} .$$

Erityisesti, jos Σ=P, $\rho(X_i,F_j)=a_{ij}$ eli **A** sisältää muuttujien ja faktorien väliset korrelaatiokertoimet.

Käytännössä faktorianalyysi tehdään otoskorrelaatiomatriisista **R** lähtien eli sille joudutaan hakemaan perusyhtälön

$$\mathbf{R} = \mathbf{AA}' + \Psi^2$$

mukainen esitys, kun yhteisfaktorit oletetaan korreloimattomiksi.

(Standardoidun) muuttujan X_i varianssi voidaan esittää tällöin muodossa

$$\mathrm{var}(X_i) = 1 = a_{i1}^2 + a_{i2}^2 + ... + a_{ir}^2 + \psi_i^2 , \; i = 1,2,...,p .$$

Faktorilatausten neliösummaa sanotaan muuttujan X_i *kommunaliteetiksi* ja sitä merkitään

$$h_i^2 = a_{i1}^2 + a_{i2}^2 + ... + a_{ir}^2 = 1 - \psi_i^2 , \; i = 1,2,...., p .$$

Kommunaliteetit kuvastavat systemaattista, faktorien avulla selitettävää osaa muuttujien varianssista. On näin luonnollista sanoa muuttujavektoria

$$\mathbf{Y} = \mathbf{X} - \mu - \mathbf{U} = \mathbf{AF}$$

muuttujien **X** systemaattiseksi osaksi. Tällöin siis

$$Y_i = a_{i1}F_1 + a_{i2}F_2 + \ldots a_{ir}F_r\,,\ \ i = 1,2,\ldots,p$$

ja

$$\rho(X_i,Y_i) = \frac{h_i^2}{1 \cdot h_i} = h_i\,.$$

On tapana puhua r-ulotteisesta faktoriavaruudesta, jonka virittävät (ortogonaaliset) faktorit $F_1,F_2,\ldots,F_r$. Tällöin muuttujien **X** systemaattiset osat **Y** voidaan esittää pisteinä tai origosta lähtevinä vektoreina, joiden päätepisteet ovat

$$\mathbf{x}_i = (a_{i1},a_{i2},\ldots,a_{ir})\,,\ \ i = 1,2,\ldots,p$$

tässä faktoriavaruudessa.

Näiden vektorien pituudet ovat $\|\mathbf{x}_i\|=h_i$ ja niiden välisille kulmille ϕ_{ij} pätee

$$\cos\phi_{ij} = \cos(\mathbf{x}_i,\mathbf{x}_j) = \frac{\mathbf{x}_i'\mathbf{x}_j}{\|\mathbf{x}_i\|\cdot\|\mathbf{x}_j\|} = \frac{a_{i1}a_{j1} + \ldots + a_{ir}a_{jr}}{h_i \cdot h_j} = \frac{r_{ij}}{h_i \cdot h_j}\,,$$

sillä

$$a_{i1}a_{j1} + \ldots + a_{ir}a_{jr} = r_{ij}\,,\ \text{kun}\ i \neq j$$

perusyhtälön $\mathbf{R}=\mathbf{AA}'+\Psi^2$ mukaan.

Korrelaatiokerroin r_{ij} ei siis suoraan vastaa vektorien $\mathbf{x}_i$ ja $\mathbf{x}_j$ välisen kulman kosinia vaan

$$r_{ij} = h_i h_j \cos\phi_{ij}\,.$$

Sen sijaan todetaan helposti, että systemaattisten osien Y_i ja Y_j korrelaatiokerroin on

$$\rho(Y_i,Y_j) = \frac{a_{i1}a_{j1} + \ldots + a_{ir}a_{jr}}{h_i \cdot h_j} = \cos\phi_{ij}\,.$$

Tämän tarkastelun perusteella on oikeutettua ajatella faktorianalyysin tulosta kuvattavaksi geometrisesti siten, että muuttujaa X_i vastaa piste (tai vektori) $\mathbf{x}_i=(a_{i1},a_{i2},\ldots,a_{ir})$ r-ulotteisessa faktoriavaruudessa ja eri pisteiden väliset suhteet tulkitaan tavanomaisen euklidisen metriikan mukaisesti.

Huomattakoon lisäksi, että yhtälöstä

$$r_{ij} = h_i h_j \cos\phi_{ij}$$

saadaan kommunaliteetille h_i^2 arvio

$$r_{ij}^2 = h_i^2 h_j^2 \cos^2\phi_{ij} \le h_i^2\,.$$

On osoitettavissa vielä vahvempi tulos eli

$$R^2_{i\,.\,1,2,\ldots,i\text{-}1,\,i\text{+}1,\ldots,p} \le h_i^2\,.$$

Faktorianalyysin teko käytännössä on tuottanut tutkijoille aina hankaluuksia, koska oikean faktoriluvun r lisäksi on kyettävä määräämään muuttujien systemaattisen vaihtelun osuus eli kommunaliteetit. Faktorianalyysin historia tuntee lukuisia faktorointimenetelmiä, joita on esitelty ehkä laajimmin teoksessa H.H.Harman: *Modern Factor Analysis* (Second edition, 1967). Suomalaiselle

lukijalle on kiintoisaa tutustua myös Toivo Vahervuon ja Yrjö Ahmavaaran oppikirjaan *Johdatus faktorianalyysiin* (1958).

Ennen 1960-lukua esitetyt keinot (kuten esim. sentroidimenetelmä) on kehitetty ottamalla huomioon sen ajan rajoittuneet laskentamahdollisuudet. Tarkemmat laskentamenetelmät kuten pääakselifaktorointi ja suurimman uskottavuuden menetelmä ovat täysin syrjäyttäneet ne. Samoin on tapahtunut rotaatiomenetelmien puolella. Kuitenkin ns. graafisella rotaatiolla, jonka Survossa voi hoitaa tietokoneavusteisesti, on edelleenkin suuri merkitys taitavien tutkijoiden sovelluksissa.

5.2 Pääakselifaktorointi

Jo nimestä on pääteltävissä, että kysymyksessä on pääkomponenttianalyysin muunnos. Alkuperäisten standardoitujen muuttujien **X** asemasta tarkastellaan niiden systemaattisia osia **Y**. Tällöin

$$\mathrm{cov}(\mathbf{Y}) = \mathrm{E}(\mathbf{AFF'A'}) = \mathbf{AA'} = \mathbf{R} - \Psi^2 .$$

Olettakaamme kommunaliteetit $h_i^2=1-\psi_i^2$ tunnetuiksi, jolloin

$$\mathbf{R} - \Psi^2 = \begin{bmatrix} h_1^2 & r_{12} & \ldots & r_{1p} \\ r_{12} & h_2^2 & \ldots & r_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ r_{1p} & r_{2p} & & h_p^2 \end{bmatrix}$$

tunnetaan. Tällöin $p\times r$-faktorimatriisi **A** voidaan määrätä "pienimmän neliösumman keinolla" siten, että lauseke

$\|\mathbf{R} - \Psi^2 - \mathbf{AA'}\|^2$

minimoidaan matriisin **A** suhteen eli $\mathbf{A}=P^{(r)}(\mathbf{R}\text{-}\Psi^2)$. Pääakselifaktorointi on siis sama kuin systemaattisten osien $Y_1,Y_2,\ldots,Y_p$ pääkomponenttianalyysi.

Koska käytännössä kommunaliteetteja ei kuitenkaan tunneta, ne joudutaan korvaamaan sopivilla arvioilla. Yleisesti käytettyjä arvoja ovat:

1) $h_i^2 = \max_{i\neq j} |r_{ij}|$,

2) $h_i^2 = R^2_{i.\,1,2,\ldots,i\text{-}1,i+1,\ldots,p}$,

3) $h_i^2 = 1$.

Kun p on parinkymmenen luokkaa tai suurempi, on melko samantekevää, miten kommunaliteetit arvioidaan. Lopulliset arviot saadaan pääakseliratkaisun jälkeen laskemalla ne estimoidusta faktorimatriisista **A** määritelmän mukaisesti kaavalla

$$h_i^2 = a_{i1}^2 + a_{i2}^2 + \ldots + a_{ir}^2 , \; i = 1,2,\ldots,p .$$

Ainakin pienillä muuttujamäärillä on paras määrätä kommunaliteetit iteroimalla eli syöttämällä estimoidusta matriisista **A** lasketut kommunaliteetit ta-

kaisin korrelaatiomatriisin lävistäjälle ja toistamalla tätä menettelyä, kunnes kommunaliteetit tulevat riittävän vakaiksi. Käytännössä riittää muutama iteraatio. Tämä menettely vastaa Harmanin Minres-ratkaisua. Survossa iteratiivinen ratkaisu tapahtuu havainnollisesti aktivoimalla matriisiketju `MATRUN PFACT` toistuvasti tai tehokkaammin `FACTA`-operaatiolla, kun täsmennyksenä on `METHOD=ULS` (unweighted least squares).

5.3 Suurimman uskottavuuden faktorointi

Olkoon

$\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(N)}$

otos multinormaalijakaumasta $N(\mu,\Sigma)$, missä kovarianssimatriisi Σ toteuttaa faktorianalyysin perusyhtälön $\Sigma=\mathbf{AA}'+\Psi^2$. Tässä $\mathbf{A}$ on $p\times r$-faktorimatriisi ja Ψ^2 $p\times p$-lävistäjämatriisi.

Suurimman uskottavuuden funktion logaritmi, kun se on maksimoitu parametrin μ suhteen on tässä tapauksessa

$$\log L(\mu,\Sigma) = -\tfrac{1}{2}[pN\log(2\pi) + N\log|\Sigma| + \operatorname{tr}(\Sigma^{-1}\mathbf{M})] ,$$

missä $\mathbf{M}$ on otoksesta laskettu momenttimatriisi. Ko. lauseke on nyt maksimoitava, kun $\Sigma=\mathbf{AA}'+\Psi^2$, parametrien $\mathbf{A}$ ja Ψ suhteen.

Maksimointitehtävän luonne riippuu faktorien lukumäärästä r suhteessa muuttujien lukumäärään p. Olennaisia parametreja matriisissa Σ on $p(p+1)/2$ kappaletta ja matriisissa $\mathbf{AA}'+\Psi^2$ vastaavasti $pr+p-r(r-1)/2$. Vähennys $r(r-1)/2$ jälkimmäisessä johtuu rotaatiomahdollisuudesta.

Jos siis $p(p+1)/2 \le pr+p-r(r-1)/2$ eli $p+r \ge (p-r)^2$, vaatimus $\Sigma=\mathbf{AA}'+\Psi^2$ ei lainkaan rajoita kovarianssimatriisia Σ ja päädytään näin pääakseliratkaisuun $\mathbf{A}=P^{(r)}(\mathbf{R})$, $\Psi^2=\mathbf{0}$. Tämä tilanne on kuitenkin käytännössä harvinainen, sillä se edellyttää suhteellisesti hyvin suurta faktorilukua. Esim. tapauksessa p=20, $r\ge$15.

Jos $p+r < (p-r)^2$, mikä on siis normaali tilanne, *D.N.Lawley* (1940) on osoittanut, että ehdolla

$\mathbf{A}'\Psi^{-2}\mathbf{A}$ = lävistäjämatriisi,

joka yksikäsitteistää tuloksen kiinnittämällä rotaation, ratkaisuksi saadaan

$$\mathbf{A} = P^{(r)}(\Psi^{-1}(\mathbf{R} - \Psi^2)\Psi^{-1}) .$$

Tällä ratkaisulla on sekin etu puolellaan verrattuna pääakseliratkaisuun, että se on invariantti muuttujien $\mathbf{X}$ säännöllisissä lineaarimuunnoksissa. Tehtävä edellyttää kuitenkin iterointia, joka kauan tuotti suuria vaikeuksia, koska suppeneminen kohti lopullista ratkaisua oli kohtuuttoman hidasta. *K.G.Jöreskog* esitti kuitenkin väitöskirjassaan (1963) menettelyn, joka huomattavasti nopeuttaa konvergenssia ja tekee suurimman uskottavuuden ratkaisun käytännössä mahdolliseksi. Jöreskog on sittemmin soveltanut samaa tekniikkaa me-

nestyksellisesti faktorianalyysin yleistyksissä eli ns. rakenneyhtälömalleissa (LISREL).

Survossa suurimman uskottavuden ratkaisu saadaan `FACTA`-operaatiolla ilman `METHOD`-täsmennystä (tai täsmennyksellä `METHOD=ML`).

5.4 Rotaatiomenetelmät

Kuten edellä on käynyt ilmi, faktorianalyysin tulos ei ole yksikäsitteinen faktorimatriisin **A** osalta. Faktoreihin **F** voidaan kohdistaa säännöllinen muunnos $\mathbf{F}^{*}=\mathbf{T}'\mathbf{F}$ ilman, että faktorianalyysin malliin liittyvät perusolettamukset ja esim. perusyhtälö siitä järkkyvät. Faktorimatriisin **A** tasolla tämä rotaatiomahdollisuus tarkoittaa siirtymistä ns. rotatoituun faktorimatriisiin $\mathbf{A}^{*}=\mathbf{A}(\mathbf{T}^{-1})'$, kuten nähtiin tämän luvun alussa.

Rotaation tuottama epämääräisyys on lupa kääntää tutkijan ja tutkimuskohteen eduksi. On täysin sallittua valita sellainen rotaatiomatriisi, joka antaa mahdollisimman selkeän käsityksen muuttujien ja faktorien välisistä riippuvuuksista. Toistamalla tutkimus esim. uusissa olosuhteissa etsitään vahvistusta sille, että saavutetut tulokset ovat invariantteja.

Valittaessa rotaatioratkaisua pyritään yleensä faktorirakenteen yksinkertaisuuteen (simple structure), jonka faktorianalyysin uranuurtaja *L.L.Thurstone* on määritellyt jo 1930-luvulla. Hänen alkuperäiset vaatimuksensa olivat seuraavat:

1. Faktorimatriisin jokaisella rivillä tulee olla ainakin yksi nolla.
2. Jokaisessa sarakkeessa tulee olla ainakin *r* nollaa.
3. Jokaista faktoriparia kohti tulee olla ainakin *r* muuttujaa, joiden lataukset ovat nollia ensimmäisessä mutta ei toisessa.

Tässä "nollalla" tarkoitetaan lähellä nollaa olevia arvoja. Thurstone itse on myöhemmin vielä laajentanut näitä "yksinkertaisen rakenteen" vaatimuksia, joita ei tässä toisteta. Pikemmin on syytä pelkistäen todeta, että rotaation avulla faktorimatriisi yritetään saattaa muotoon, jossa on mahdollisimman paljon lähellä nollaa olevia latauksia, toisaalta jonkin verran itseisarvoltaan suuria latauksia ja mahdollisimman vähän "keskikokoisia". Tällainen tilanne luo edellytykset tulosten tulkinnalle, jota esim. verrataan tutkittavasta ilmiöstä esitettyihin teorioihin.

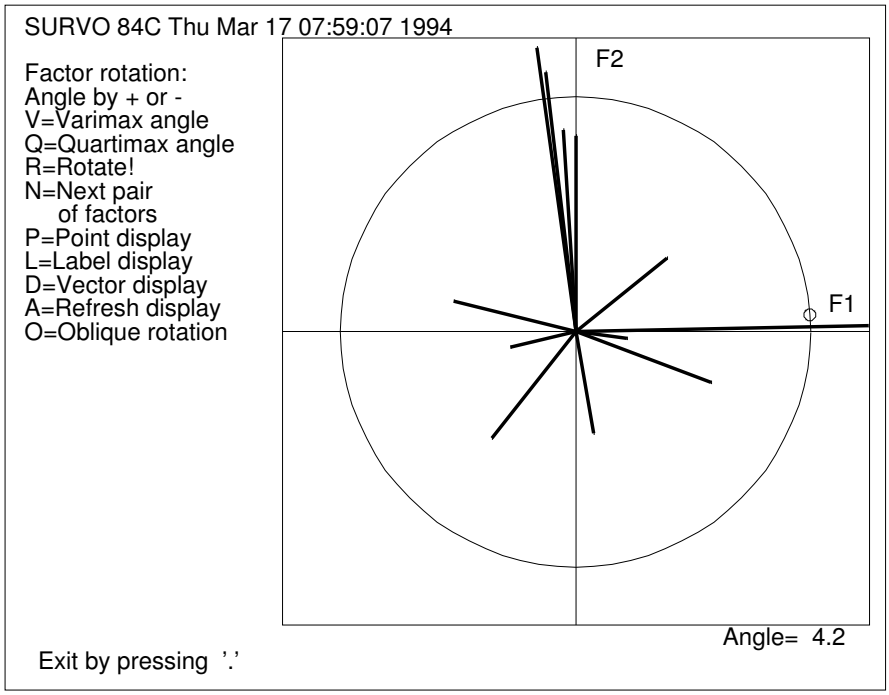
On tapana puhua "ortogonaalisesta" rotaatiosta, kun rotaatiomatriisi on ortogonaalinen. Tämä merkitsee sitä, että faktorit jäävät rotaation jälkeenkin korreloimattomiksi. Toinen vaihtoehto on "vino" rotaatio, jossa rotaatiomatriisi ei ole enää ortogonaalinen ja faktoreiden sallitaan korreloida keskenään. Kummassakin tapauksessa nykyisin haetaan ratkaisu yleensä jonkin analyyttisen kriteerin perusteella, joka tekee ratkaisusta objektiivisemman. Silti taitavissa käsissä graafisella rotaatiolla on edelleen oma paikkansa käytännön sovelluksissa.

5.4.1 Graafinen rotaatio

Aikaisemmin yleistä oli suorittaa rotaatio graafisesti piirtämällä faktoriavaruuden kaksiulotteisia projektioita eri faktoriparien suhteen, valita kussakin tarkoituksenmukainen 2-ulotteinen rotaatio, päivittää faktorimatriisi valitun rotaation mukaisesti ja jatkaa uudella faktoriparilla niin kauan, kun saavutetaan toivotut ehdot täyttävä ratkaisu.

On helppo ymmärtää, että graafinen rotaatio syrjäytyi täysin, kun tuli mahdolliseksi käyttää analyyttisia kriteereitä, jotka voi ohjelmoida tietokoneelle. Vuorovaikutteista graafista ratkaisua ei pitkään voinut edes ajatella ohjelmoitavaksi. Survossa graafinen rotaatio on ollut kuitenkin saatavilla jo SURVO 76 -versiossa vuodesta 1977 lähtien, koska järjestelmä jo silloin oli luonteeltaan vuorovaikutteinen.

Nykyisessä Survossa (SURVO 84C) kaikki rotaatioratkaisut tehdään `ROTATE`-operaatiolla ja graafiseen rotaatioon päästään antamalla täsmennys `ROTATION=GRAPHICAL` . Rotaatio alkaa tällöin kahden ensimmäisen faktorin projektiosta ja käyttäjä voi ohjata rotaatiota kuvaruudulla ilmoitetuilla napeilla. Perustilanteena on ortogonaalinen rotaatio, jolloin käyttäjän riittää ohjata kohdistin osoittamaan uutta X-akselin paikkaa, joka näin määrää kiertokulman. Yksittäisen 2-ulotteisen rotaation jälkeen voi vaihtaa seuraavaan faktoripariin ja jatkaa uudella kulman valinnalla. On mahdollista myös siirtyä vinoon rotaatioon, jolloin käyttäjä joka vaiheessa joutuu osoittamaan sekä uuden X- että uuden Y-akselin paikat. Tässä tekstissä on turha selittää kaikkia yksityiskohtia. Menettelyn oppii vain kokeilemalla käytännössä. Tyypillinen näkymä graafista rotaatiota aloitettaessa on seuraavanlainen:



Tähän on päästy seuraavien vaiheiden kautta. Ensin on laskettu aineiston KYMMEN 12 muuttujasta (lajimuuttujat sekä pituus että paino) korrelaatiot ja sovellettu saatuun korrelaatiomatriisiin FACTA-operaatiota 3 faktorilla. Suurimman uskottavuuden faktorimatriisi tallentuu matriisitiedostoon FACT.M . Aktivoimalla ROTATE-komento (rivillä 21) päästään edellä esitettyyn graafisen rotaation alkutilanteeseen.

```
 22  1 SURVO 84C EDITOR Thu Mar 17 07:58:45 1994          D:\M\MEN\ 100 100 0
  1 *
  2 *MASK=--AAAAAAAAAAA-----
  3 *CORR KYMMEN
  4 *FACTA CORR.M,3,CUR+1_
  5 *Factor analysis: Maximum Likelihood (ML) solution
  6 *Factor matrix
  7 *             F1     F2     F3    h^2
  8 *M100       0.997  0.022  0.001  0.995
  9 *Pituush    0.174 -0.021 -0.088  0.038
 10 *Kuula     -0.043  0.689 -0.405  0.641
 11 *Korkeus   -0.414  0.105 -0.276  0.259
 12 *M400       0.461 -0.173  0.505  0.497
 13 *Aidat      0.312  0.251  0.107  0.172
 14 *Kiekko     0.000  0.668 -0.519  0.715
 15 *Seiväs     0.063 -0.344 -0.096  0.131
 16 *Keihäs    -0.220 -0.051 -0.183  0.085
 17 *M1500     -0.285 -0.363  0.647  0.631
 18 *Pituus    -0.131  0.967  0.119  0.967
 19 *Paino     -0.102  0.886 -0.156  0.820
 20 *
 21 *ROTATE FACT.M,3,CUR+1_ / ROTATION=GRAPHICAL MODE=VGA
```

5.4.2 Analyyttiset rotaatiomenetelmät

Analyyttisissa menetelmissä käytetään yksinkertaista rakennetta jollain tavoin kuvaavaa, rotatoidusta faktorimatriisista $\mathbf{A}^*=\mathbf{A}(\mathbf{T}^{-1})'$ riippuvaa mittaa $m(\mathbf{A}^*)$, joka maksimoidaan (tai minimoidaan) rotaatiomatriisin $\mathbf{T}$ suhteen. Näiden mittojen tyypillinen ominaisuus on se, että ne saavuttavat optimiarvonsa silloin, kun matriisin $\mathbf{A}^*$ latauksset ovat mahdollisimman vaihtelevia.

Houkuttelevaa olisi ajatella yksinkertaisesti latausten varianssin maksimointia, mutta rotaatiomatriisin ollessa ortogonaalinen, on

$$\|\mathbf{A}^*\|^2 = \|\mathbf{AT}\|^2 = \mathrm{tr}(\mathbf{ATT'A'}) = \mathrm{tr}(\mathbf{AA'}) = \|\mathbf{A}\|^2 = \sum_{i=1}^{p} h_i^2$$

eli latausten neliösumma ja kommunaliteettien summa säilyy tällaisessa muunnoksessa. Sen vuoksi funktio $m()$ on yleensä neljättä astetta.

Ortogonaalisten rotaatioiden puolella tunnetaan parhaiten **Quartimax**- ja **Varimax**-menetelmät. Quartimax-menetelmässä yksinkertaisesti maksimoidaan latausten neljänsien potenssien summa

$$m(\mathbf{A}^*) = \sum_{i=1}^{p} \sum_{j=1}^{r} a_{ij}^{*4}.$$

Varimax-rotaatiossa (*H.F.Kaiser* 1958) maksimoidaan latausten neliöitten

sarakkeittain (faktoreittain) laskettujen varianssien summa eli

$$m(\mathbf{A}^*) = p^2 \sum_{j=1}^{r} v_j^2 ,$$

missä

$$v_j^2 = \frac{1}{p} \sum_{i=1}^{p} [\, a_{ij}^{*2} - \frac{1}{p} \sum_{k=1}^{p} a_{kj}^{*2}]^2$$

$$= \frac{1}{p} \sum_{i=1}^{p} a_{ij}^{*4} - \frac{1}{p^2} [\, \sum_{i=1}^{p} a_{ij}^{*2}]^2 .$$

Tavallisesti vielä poistetaan muuttujien väliset kommunaliteettierot jakamalla lataukset riveittäin luvuilla h_i ja kertomalla ne takaisin samoilla luvuilla rotaation jälkeen.

Käytännössä sekä Quartimax- että Varimax-ratkaisu joudutaan etsimään iteratiivisesti tekemällä peräkkäisiä kaksiulotteisia kiertoja, kuten tapahtuu graafisessakin rotaatiossa. Faktoriparit valitaan systemaattisesti. Esim. 4 faktorin tapauksessa valintajärjestys voi olla (1,2) (1,3) (1,4) (2,3) (2,4) (3,4) (1,2),... Vaikka optimaalinen kiertokulma on suhteellisen yksinkertainen lauseke kummassakin ratkaisussa, kunkin kierron yhteydessä kuitenkin faktorimatriisin lataukset muuttuvat aina kahden faktorin osalta, jolloin joudutaan toistuvasti palaamaan uudelleen samoihin akselipareihin. Kiertokulmat suppenevat tavallisesti melko nopeasti kohti nollaa ja iterointi keskeytyy, kun riittävä tarkkuus on saavutettu.

Survossa Varimax-rotaatio valitaan `ROTATE`-operaatiossa automaattisesti, ellei `ROTATION`-täsmennyksellä ole toisin määrätty. Quartimax-menetelmä ei ole suoraan mukana, mutta se samoin kuin Varimax on käytettävissä apuna graafisessa rotaatiossa (`ROTATION=GRAPHICAL`), jolloin tutkija saa napilla Q näkyville Quartimax-menetelmän mukaisen parhaan kiertokulman ja vastaavasti napilla V Varimax-ehdotuksen. Näin kumpaakin menetelmää voi soveltaa askeltaen graafisen rotaation yhteydessä.

5.4.3 Vinot rotaatiot

Kun faktorianalyysin lähtökohtana on muuttujien **X** korrelaatiomatriisi **R**, on luonnollista edellyttää, että faktorit normeerataan siten, että niiden varianssit ovat ykkösiä, jolloin Φ on faktorien korrelaatiomatriisi. Jos alkuperäinen faktorirakenne on ortogonaalinen, rotaatiossa $\mathbf{A}^*=\mathbf{A}(\mathbf{T}^{-1})'$perusyhtälö

$$\mathbf{\Sigma}=\mathbf{A}\mathbf{A}'+\Psi^2$$

saa muodon

$$\mathbf{\Sigma}=\mathbf{A}^*\mathbf{T}'\mathbf{T}\mathbf{A}^{*\prime}+\Psi^2 ,$$

jolloin

$$\mathbf{T}'\mathbf{T} = \Phi .$$

Koska korrelaatiomatriisin Φ lävistäjäalkiot ovat ykkösiä, seuraa yhtälöstä $\mathbf{T}'\mathbf{T}=\Phi$, että matriisin $\mathbf{T}$ sarakkeet ovat yksikkövektorin mittaisia. Nämä sarakkeet koostuvat faktoriavaruuden uusien vinojen akselien suuntakosineista alkuperäisten ortogonaalisten akselien suhteen. Tällöin matriisi

$$\mathbf{S} = \mathbf{AT} ,$$

jota sanotaan vinon rotaation rakennematriisiksi (oblique structure), antaa muuttujien (systemaattisten osien) projektiot uusilla vinokulmaisilla koordinaattiakseleilla. Varsinainen rotatoitu faktorimatriisi $\mathbf{A}^*$ (pattern matrix), joka antaa vinokulmaiset koordinaatit, on tällöin

$$\mathbf{A}^* = \mathbf{A}(\mathbf{T}^{-1})' = \mathbf{S}\mathbf{T}^{-1}(\mathbf{T}^{-1})' = \mathbf{S}\Phi^{-1}$$

eli kääntäen

$$\mathbf{S} = \mathbf{A}^*\Phi .$$

Matriisi $\mathbf{S}$ on myös muuttujien $\mathbf{X}$ ja rotatoitujen faktoreiden $\mathbf{F}$ välisten korrelaatiokertoimien matriisi, sillä

$$\rho(\mathbf{X},\mathbf{F}) = \mathrm{cov}(\mathbf{X},\mathbf{F}) = \mathrm{E}[(\mathbf{A}^*\mathbf{F} + \mathbf{U})\mathbf{F}'] = \mathbf{A}^*\Phi = \mathbf{S} .$$

Survon vinoissa rotaatioratkaisuissa ko. matriiseita vastaavat seuraavat matriisitiedostot:

Matriisi	Tiedosto
$\mathbf{A}$	`FACT.M` tai `PFACT.M`
$\mathbf{T}$	`TFACT.M`
$\mathbf{A}^*$	`AFACT.M`
Φ	`RFACT.M`
$\mathbf{S}$	`SFACT.M`.

Analyyttisia vinoja rotaatioratkaisuja, joita on ehdotettu lukuisia, Survossa edustaa ns. suora Oblimin-menetelmä (*Jennrich* ja *Sampson* 1966). Toisena vaihtoehtona on *Yrjö Ahmavaaran* alunperin esittämä kosiniratkaisu, jota vastaavan analyyttisen ratkaisun ovat kehittäneet *T.Markkanen*, *S.Mustonen* ja *M.Tienari* v. 1962.

Oblimin-menetelmässä (kts. Harman ss. 334-341) minimoidaan lauseke

$$\sum_{j=1}^{r} \sum_{k=1}^{j-1} [\sum_{i=1}^{p} a_{ik}^{*2} a_{ij}^{*2} - \frac{\delta}{p} \sum_{i=1}^{p} a_{ik}^{*2} \sum_{i=1}^{p} a_{ij}^{*2}]$$

rotaatiomatriisin $\mathbf{T}$ suhteen ehdolla, että $\mathrm{diag}(\mathbf{T}'\mathbf{T})=\mathbf{I}$. Minimoitava lauseke riippuu parametrista δ, jonka arvoksi yleensä valitaan 0 tai negatiivinen luku. Mitä negatiivisempi δ on, sitä ortogonaalisemmaksi ratkaisu muodostuu. Suositeltavin arvo on yksinkertaisesti $\delta=0$, joka Survossa (kun `ROTATE`-operaatiota käytetään täsmennyksellä `ROTATION=OBLIMIN`) on oletusarvona. Tässä tapauksessa on helppo nähdä lausekkeen muodosta, miten "nollat" ja "suuret" lataukset pyritään saamaan eri faktoreille yksinkertaisen rakenteen hengessä.

Ahmavaaran kosinirotaatiossa yksinkertaisen rakenteen vaatimukset saadaan yleensä voimaan valitsemalla muuttujista $\mathbf{X}$ r kappaletta ns. faktorimuuttujiksi

asettamalla rotatoidut faktoriakselit kulkemaan pitkin ao. muuttujavektoreita. Tällöin kukin faktorimuuttuja saa suuren latauksen omalla faktorillaan ja nollalatauksen muilla faktoreilla. Tähän samaan pyritään usein myös graafisessa rotaatiossa.

Kosiniratkaisun teknisenä ongelmana on ollut, miten faktorimuuttujat olisi tällöin valittava, koska vaihtoehtoja on binomikertoimen C(p,r) ilmaisema määrä. On luonnollista yrittää saada faktorimuuttujat mahdollisimman ortogonaalisiksi toistensa suhteen.

Markkasen, Mustosen ja Tienarin kehittämässä laskentamenetelmässä tätä päämäärää tavoitellaan maksimoimalla normeerattujen vektorien

$$\mathbf{x}_{i_j} / \|\mathbf{x}_{i_j}\| \ , \ j = 1,2,...,r$$

muodostaman matriisin determinantin itseisarvo valitsemalla indeksit i_1,i_2,...,i_r luvuista 1,2,...,p. Ko. determinantti vastaa vektorien virittämän suuntaissärmiön tilavuutta ja mittaa siten myös vektorien ortogonaalisuuden astetta.

Koska käytännössä on mahdotonta käydä läpi kaikkia faktorikombinaatioita, Survon ratkaisussa (`ROTATION=COS`) tyydytään yksinkertaisesti läpikäymään vain p vaihtoehtoa lähtemällä vuorollaan liikkeelle yhdestä muuttujasta, liittämällä siihen toinen, joka on ensiksi valitun suhteen mahdollisimman ortogonaalinen, tämän jälkeen kolmas, joka on mahdollisimman ortogonaalinen em. kahden muodostamaa tasoa vastaan jne. Tämä ei takaa, että aina ehdoton globaali optimiratkaisu löytyisi, mutta antaa kuitenkin optimia lähellä olevan tuloksen.

Kosinirotaatiota koskevan täsmennyksen voi antaa myös esim. muodossa `ROTATION=COS,0.4` , jossa lisäparametri `0.4` osoittaa, että faktorimuuttujiksi saa valita vain sellaisia, joiden kommunaliteetit ovat ainakin 0.4 . Näin vältetään se, etteivät kovin alhaisen yhteisvaihteluosuuden omaavat muuttujat pääse vaikuttamaan tulokseen. Oletusarvo tälle kommunaliteettien alarajalle on 0.3 .

5.4.4 Esimerkki

Tarkastelemme faktorianalyysin menetelmien toimivuutta keinotekoisessa esimerkkitilanteessa, jossa faktorimatriisi **A** ja faktorien korrelaatiomatriisi Φ tunnetaan. Näiden ja perusyhtälön $\mathbf{R}=\mathbf{A}\Phi\mathbf{A}'+\Psi^2$ avulla konstruoimme korrelaatiomatriisin **R**, jota sitten lähdemme analysoimaan aluksi puhtaasti teoreettisilla suureilla ja lopuksi luomalla tätä mallia noudattavan simuloidun otoksen, jonka analysoimme vastaavin keinoin.

Tässä ovat 8 muuttujan ja 3 faktorin faktorimatriisi **A** ja faktorien korrelaatiomatriisi Φ. Muuttujat on selvyyden vuoksi jaettu kolmeen ryhmään, X-, Y- ja Z-muuttujiin siten, että kukin ryhmä vastaa yhtä faktoria vieläpä niin, että ryhmien ensimmäiset ovat puhtaita faktorimuuttujia (lataukset muilla faktoreilla nollia).

Matriisit talletetaan matriisitiedostoiksi `A.MAT` ja `PHI.MAT`:

```
 13  1 SURVO 84C EDITOR Wed Mar 23 08:59:23 1994          D:\M\MEN\ 240 100 0
  1 *
  2 *Faktorimatriisi:
  3 *MATRIX A
  4 *///   F1    F2    F3
  5 *X1    0.8   0     0
  6 *X2    0.7   0.3  -0.2
  7 *X3   -0.6   0.5   0
  8 *X4    0.5   0.2   0.1
  9 *Y1    0     0.9   0
 10 *Y2    0.1   0.7  -0.2
 11 *Y3    0.3  -0.5   0.3
 12 *Y4    0.2   0.3   0
 13 *Z1    0     0     0.6
 14 *Z2   -0.2  -0.2   0.5
 15 *Z3    0.3   0    -0.4
 16 *Z4   -0.3   0.3   0.3
 17 *
 18 *Faktorien korrelaatiomatriisi:
 19 *MATRIX PHI
 20 *///   F1   F2   F3
 21 *F1    1    0.4   0.2
 22 *F2    0.4  1    -0.1
 23 *F3    0.2  -0.1  1
 24 *
 25 *Talletetaan matriisit A ja PHI:
 26 *MAT SAVE A
 27 *MAT SAVE PHI
 28 *
```

Matriisi **R** lasketaan perusyhtälöstä. Ominaisfaktorien varianssit Ψ^2 (tässä `PSI`) otetaan huomioon täydentämällä matriisia $\mathbf{A\Phi A}'$ yksinkertaisesti niin, että sen lävistäjäalkiot tulevat ykkösiksi.

```
  1  1 SURVO 84C EDITOR Wed Mar 23 09:05:38 1994          D:\M\MEN\ 240 100 0
 28 *
 29 *Lasketaan korrelaatiomatriisi R perusyhtälöstä R=A*PHI*A'+PSI :
 30 *MAT DIM A /* rowA=12 colA=3
 31 *MAT R=A'           / *R~A' 3*12
 32 *MAT R=PHI*R        / *R~PHI*A' 3*12
 33 *MAT R=A*R          / *R~A*PHI*A' 12*12
 34 *MAT D=VD(R)        / *D~VD(A*PHI*A') 12*1
 35 *MAT D=DV(D)        / *D~DV(VD(A*PHI*A')) D12*12
 36 *MAT I=IDN(rowA,rowA)
 37 *MAT PSI!=I-D       / *PSI~IDN-DV(VD(A*PHI*A')) D12*12
 38 *MAT R=R+PSI        / *R~A*PHI*A'+PSI 12*12
 39 *
```

Mallin mukainen korrelaatiomatriisi näyttää seuraavalta:

```
 21  1 SURVO 84C EDITOR Wed Mar 23 09:12:59 1994          D:\M\MEN\ 240 100 0
 39 *
 40 *Saatu korrelaatiomatriisi R:
 41 *LOADM R,##.###,CUR+1_
 42 *A*PHI*A'+PSI
 43 *              X1     X2     X3     X4     Y1     Y2     Y3     Y4     Z1
 44 *X1         1.000  0.624 -0.320  0.480  0.288  0.272  0.128  0.256  0.096
 45 *X2         0.624  1.000 -0.168  0.501  0.540  0.516 -0.093  0.336 -0.054
 46 *X3        -0.320 -0.168  1.000 -0.165  0.234  0.176 -0.301 -0.002 -0.102
 47 *X4         0.480  0.501 -0.165  1.000  0.351  0.297  0.039  0.237  0.108
 48 *Y1         0.288  0.540  0.234  0.351  1.000  0.684 -0.369  0.342 -0.054
 49 *Y2         0.272  0.516  0.176  0.297  0.684  1.000 -0.353  0.296 -0.150
 50 *Y3         0.128 -0.093 -0.301  0.039 -0.369 -0.353  1.000 -0.091  0.246
 51 *Y4         0.256  0.336 -0.002  0.237  0.342  0.296 -0.091  1.000  0.006
 52 *Z1         0.096 -0.054 -0.102  0.108 -0.054 -0.150  0.246  0.006  1.000
 53 *Z2        -0.144 -0.321 -0.057 -0.108 -0.297 -0.345  0.255 -0.135  0.288
 54 *Z3         0.176  0.270 -0.052  0.108  0.144  0.202 -0.116  0.092 -0.204
 55 *Z4        -0.096 -0.081  0.147 -0.009  0.135  0.051 -0.048  0.021  0.126
 56 *
 57 *              Z2     Z3     Z4
 58 *X1        -0.144  0.176 -0.096
 59 *X2        -0.321  0.270 -0.081
 60 *X3        -0.057 -0.052  0.147
 61 *X4        -0.108  0.108 -0.009
 62 *Y1        -0.297  0.144  0.135
 63 *Y2        -0.345  0.202  0.051
 64 *Y3         0.255 -0.116 -0.048
 65 *Y4        -0.135  0.092  0.021
 66 *Z1         0.288 -0.204  0.126
 67 *Z2         1.000 -0.246  0.099
 68 *Z3        -0.246  1.000 -0.120
 69 *Z4         0.099 -0.120  1.000
 70 *
```

Survon FACTA-operaatiolla lasketaan suurimman uskottavuuden ratkaisu kolmella faktorilla:

```
 16  1 SURVO 84C EDITOR Wed Mar 23 09:16:52 1994          D:\M\MEN\ 240 100 0
 70 *
 71 *Suurimman uskottavuuden ratkaisu:
 72 *FACTA R,3,CUR+1_
 73 *Factor analysis: Maximum Likelihood (ML) solution
 74 *Factor matrix
 75 *              F1     F2     F3    h^2
 76 *X1         0.556  0.574  0.027  0.640
 77 *X2         0.793  0.323 -0.102  0.744
 78 *X3         0.043 -0.602  0.073  0.370
 79 *X4         0.517  0.328  0.146  0.396
 80 *Y1         0.828 -0.309  0.169  0.810
 81 *Y2         0.744 -0.245 -0.048  0.616
 82 *Y3        -0.297  0.502  0.190  0.376
 83 *Y4         0.415  0.041  0.063  0.178
 84 *Z1        -0.091  0.230  0.547  0.360
 85 *Z2        -0.399  0.116  0.411  0.342
 86 *Z3         0.269  0.062 -0.354  0.202
 87 *Z4         0.022 -0.204  0.320  0.144
 88 *
```

Ratkaisu sellaisenaan muistuttaa vain vähäisesti alkuperäistä. Tilanne muuttuu kuitenkin täysin selkeäksi, kun sovelletaan saatuun faktorimatriisiin (FACTA on tallettanut sen matriisitiedostoon FACT.M) kosinirotaatiota (ROTATE-operaatio täsmennyksellä ROTATION=COS):

```
 55  1 SURVO 84C EDITOR Wed Mar 23 09:22:15 1994          D:\M\MEN\ 240 100 0
 88 *
 89 *Kosinirotaatio:
 90 *ROTATE FACT.M,3,CUR+1  / ROTATION=COS      RESULTS=100_
 91 *Rotated factor matrix AFACT.M=FACT.M*inv(TFACT.M)'
 92 *              F1     F2     F3 Sumsqr
 93 *X1         0.800 -0.000 -0.000  0.640
 94 *X2         0.700  0.300 -0.200  0.620
 95 *X3        -0.600  0.500  0.000  0.610
 96 *X4         0.500  0.200  0.100  0.300
 97 *Y1         0.000  0.900 -0.000  0.810
 98 *Y2         0.100  0.700 -0.200  0.540
 99 *Y3         0.300 -0.500  0.300  0.430
100 *Y4         0.200  0.300  0.000  0.130
101 *Z1         0.000  0.000  0.600  0.360
102 *Z2        -0.200 -0.200  0.500  0.330
103 *Z3         0.300  0.000 -0.400  0.250
104 *Z4        -0.300  0.300  0.300  0.270
105 *Sumsqr     2.100  2.150  1.040  5.290
106 *
107 *Rotation matrix TFACT.M
108 *              F1     F2     F3
109 *F1         0.696  0.920 -0.152
110 *F2         0.718 -0.343  0.383
111 *F3         0.034  0.187  0.911
112 *
113 *Factor correlation matrix RFACT.M
114 *              F1     F2     F3
115 *F1         1.000  0.400  0.200
116 *F2         0.400  1.000 -0.100
117 *F3         0.200 -0.100  1.000
118 *
119 *The factor structure matrix SFACT.M is obtained by the commands:
120 *MAT SFACT.M=AFACT.M*RFACT.M
121 *MAT LOAD SFACT.M,12.123,CUR+1
122 *
```

Kaikki, mitä näkyy riveillä 91-121, on ROTATE-operaation kirjoittamaa tulosta. Ratkaisu on identtinen lähtökohtana olevan faktorirakenteen kanssa.
Tässä tapauksessa myös (iteroitu) pääakselifaktorointi (FACTA-operaatiossa METHOD=ULS) tuottaa tuloksen, josta seuraa kosinirotaatiolla "oikea" ratkaisu.

Muilla automaattisilla rotaatiomenetelmillä ei löydetä alkuperäistä **A**-matriisia. Esim. Oblimin-menetelmällä saataisiin:

```
 42  1 SURVO 84C EDITOR Wed Mar 23 09:40:45 1994          D:\M\MEN\ 240 100 0
 88 *
 89 *ROTATE FACT.M,3,CUR+1  / ROTATION=OBLIMIN_
 90 *Rotated factor matrix AFACT.M=FACT.M*inv(TFACT.M)'
 91 *              F1     F2     F3 Sumsqr
 92 *X1         0.721  0.376 -0.015  0.661
 93 *X2         0.803  0.096 -0.201  0.694
 94 *X3        -0.162 -0.594  0.023  0.380
 95 *X4         0.630  0.132  0.092  0.423
 96 *Y1         0.683 -0.562  0.020  0.783
 97 *Y2         0.565 -0.436 -0.180  0.542
 98 *Y3        -0.025  0.521  0.274  0.347
 99 *Y4         0.408 -0.094  0.003  0.175
100 *Z1         0.167  0.136  0.582  0.385
101 *Z2        -0.193  0.144  0.484  0.292
102 *Z3         0.159  0.050 -0.394  0.183
103 *Z4         0.041 -0.260  0.303  0.161
104 *Sumsqr     2.632  1.417  0.977  5.026
105 *
```

Tämä on kaukana lähtökohdasta, mutta ei pidä mennä oikopäätä tuomitsemaan menetelmää kosinirotaatiota huonommaksi. Puhtaiden faktorimuuttujien olemassaolo suosii kosiniratkaisua.

Edellä käsiteltiin täysin "tarkkaa" korrelaatiomatriisia. Tutkitaan, mitä tapahtuu, kun kohteena on 300 havainnon otos multinormaalijakaumasta, jolla on tämä korrelaatiomatriisi. Luodaan otos:

```
 45  1 SURVO 84C EDITOR Wed Mar 23 13:25:50 1994          D:\M\MEN\ 240 100 0
123 *.....................................................................
124 *MAT LOAD R,CUR+1
125 *MATRIX MSN
126 *///  Keski Hajonta
127 *X1     0     1
128 *X2     0     1
129 *X3     0     1
130 *X4     0     1
131 *Y1     0     1
132 *Y2     0     1
133 *Y3     0     1
134 *Y4     0     1
135 *Z1     0     1
136 *Z2     0     1
137 *Z3     0     1
138 *Z4     0     1
139 *
140 *MAT SAVE MSN
141 */MNSIMUL R,MSN,XYZ,300  / RND=rand(23031994)_
142 *
```

Lasketaan otoskorrelaatiokertoimet, faktoroidaan ja rotatoidaan kosinimenetelmällä:

```
 53  1 SURVO 84C EDITOR Wed Mar 23 13:29:31 1994          D:\M\MEN\ 240 100 0
142 *CORR XYZ
143 *FACTA CORR.M,3
144 *ROTATE FACT.M,3,CUR+1 / ROTATION=COS     RESULTS=100_
145 *Rotated factor matrix AFACT.M=FACT.M*inv(TFACT.M)'
146 *             F1     F2     F3 Sumsqr
147 *X1         0.816  0.000 -0.000  0.665
148 *X2         0.620  0.393 -0.237  0.595
149 *X3        -0.601  0.570 -0.076  0.692
150 *X4         0.583  0.159  0.081  0.371
151 *Y1        -0.000  0.897 -0.000  0.805
152 *Y2         0.042  0.737 -0.232  0.598
153 *Y3         0.335 -0.449  0.350  0.436
154 *Y4         0.197  0.321 -0.140  0.161
155 *Z1         0.000 -0.000  0.651  0.424
156 *Z2        -0.130 -0.187  0.587  0.397
157 *Z3         0.323 -0.069 -0.514  0.373
158 *Z4        -0.259  0.222  0.256  0.182
159 *Sumsqr     2.091  2.247  1.363  5.700
160 *
161 *Rotation matrix TFACT.M
162 *             F1     F2     F3
163 *F1         0.666  0.933 -0.182
164 *F2         0.741 -0.234  0.576
165 *F3        -0.091  0.273  0.797
166 *
167 *Factor correlation matrix RFACT.M
168 *             F1     F2     F3
169 *F1         1.000  0.423  0.233
170 *F2         0.423  1.000 -0.087
171 *F3         0.233 -0.087  1.000
172 *
```

Havaitaan, että ainakin tässä tapauksessa tulokset vastaavat hyvin odotettuja. Pian esiteltävän ns. transformaatioanalyysin avulla on mahdollista tutkia vastaavuutta paremmin kuin tässä on tehtävissä silmämääräisesti.

Vastaava Oblimin-ratkaisu on seuraava:

```
 41  1 SURVO 84C EDITOR Wed Mar 23 13:31:01 1994          D:\M\MEN\ 240 100 0
144 *ROTATE FACT.M,3,CUR+1 / ROTATION=OBLIMIN_
145 *Rotated factor matrix AFACT.M=FACT.M*inv(TFACT.M)'
146 *              F1     F2     F3 Sumsqr
147 *X1         0.698  0.455 -0.026  0.695
148 *X2         0.819  0.054 -0.172  0.704
149 *X3        -0.043 -0.661  0.043  0.440
150 *X4         0.649  0.271  0.081  0.502
151 *Y1         0.764 -0.469  0.146  0.825
152 *Y2         0.619 -0.445 -0.092  0.591
153 *Y3        -0.029  0.548  0.234  0.356
154 *Y4         0.415 -0.109 -0.081  0.190
155 *Z1         0.124  0.235  0.591  0.420
156 *Z2        -0.159  0.238  0.507  0.339
157 *Z3         0.120  0.030 -0.488  0.253
158 *Z4         0.016 -0.169  0.277  0.105
159 *Sumsqr     2.777  1.592  1.050  5.419
160 *
```

5.5 Faktoripistemäärät

Faktoroinnin ja rotaation jälkeen halutaan usein tietää myös faktorien arvot havaintokohtaisesti. Näiden faktoripistemäärien laskemiseen ei ole yksikäsitteistä tapaa, koska faktorianalyysin mallia

$$\mathbf{X} = \mu + \mathbf{AF} + \mathbf{U}$$

vastaavia yhtälöitä ei voi sellaisenaan ratkaista faktorien $\mathbf{F}$ suhteen.

Paras tapa selvittää asia lienee eräänlaisen regressiomenetelmän soveltaminen. Etsitään faktoreille $\mathbf{F}$ muotoa $\mathbf{K}(\mathbf{X} - \mu)$ oleva likimääräinen esitys pienimmän neliösumman keinolla minimoimalla

$$f(\mathbf{K}) = \mathrm{E}(\|\mathbf{F} - \mathbf{K}(\mathbf{X} - \mu)\|^2) = \mathrm{E}[\mathrm{tr}(\mathbf{F} - \mathbf{K}(\mathbf{X} - \mu))'(\mathbf{F} - \mathbf{K}(\mathbf{X} - \mu))]$$

$r \times p$-matriisin $\mathbf{K}$ suhteen. Rajoitumme ortogonaalisten faktorien tapaukseen, jolloin $\mathrm{E}(\mathbf{FF}')=\mathbf{I}$. Tällöin minimi saavutetaan, kun $\mathbf{K}=\mathbf{A}'\mathbf{R}^{-1}$.

Väitteen todistamiseksi kehitellään lauseketta $f(\mathbf{K})$ seuraavasti:

$$f(\mathbf{K}) = \mathrm{E}\,\mathrm{tr}[(\mathbf{F} - \mathbf{K}(\mathbf{X} - \mu))(\mathbf{F} - \mathbf{K}(\mathbf{X} - \mu))']$$

$$= \mathrm{E}[\mathrm{tr}(\mathbf{FF}' - \mathbf{F}(\mathbf{X} - \mu)'\mathbf{K}' - \mathbf{K}(\mathbf{X} - \mu)\mathbf{F} + \mathbf{K}(\mathbf{X} - \mu)(\mathbf{X} - \mu)'\mathbf{K}')]$$

$$= \mathrm{tr}(\mathbf{I} - \mathbf{A}'\mathbf{K}' - \mathbf{KA} + \mathbf{KRK}') .$$

Käyttämällä hyväksi korrelaatiomatriisin $\mathbf{R}$ Cholesky-hajotelmaa $\mathbf{R}=\mathbf{CC}'$ osoitamme nyt, että

$$\Delta = f(\mathbf{K}) - f(\mathbf{A}'\mathbf{R}^{-1}) \geq 0 .$$

$$\Delta = \mathrm{tr}(\mathbf{I} - 2\mathbf{KA} + \mathbf{KRK}') - \mathrm{tr}(\mathbf{I} - 2\mathbf{A}'\mathbf{R}^{-1}\mathbf{A} + \mathbf{A}'\mathbf{R}^{-1}\mathbf{R}\mathbf{R}^{-1}\mathbf{A})$$

$$= \mathrm{tr}(\mathbf{KRK}' - 2\mathbf{KA} + \mathbf{A}'\mathbf{R}^{-1}\mathbf{A})$$

$$= \mathrm{tr}[(\mathbf{KC})(\mathbf{C}'\mathbf{K}') - 2(\mathbf{KC})(\mathbf{C}^{-1}\mathbf{A}) + (\mathbf{A}'\mathbf{C}'^{-1})(\mathbf{C}^{-1}\mathbf{A})]$$

$$= \mathrm{tr}(\mathbf{LL}' - 2\mathbf{LB}' + \mathbf{BB}') ,$$

missä $\mathbf{L}=\mathbf{KC}$ ja $\mathbf{B}'=\mathbf{C}^{-1}\mathbf{A}$. Tällöin

$$\Delta = \mathrm{tr}(\mathbf{LL}' - \mathbf{LB}' - \mathbf{BL}' + \mathbf{BB}') = \mathrm{tr}(\mathbf{L} - \mathbf{B})(\mathbf{L} - \mathbf{B})' .$$

Nähdään siis, että $\Delta \geq 0$ ja $\Delta = 0$ vain, kun $\mathbf{L}=\mathbf{B}$ eli kun $\mathbf{K}=\mathbf{A}'\mathbf{R}^{-1}$.

Matriisin **R** kääntäminen voi olla ongelmallista ja se voidaankin välttää seuraavalla keinolla, jonka on esittänyt *W.Ledermann* v.1938.

Perusyhtälön $\mathbf{R}=\mathbf{AA}'+\Psi^2$ nojalla on

$$\mathbf{A}'\Psi^{-2}\mathbf{R} = \mathbf{A}'\Psi^{-2}(\mathbf{AA}' + \Psi^2) = (\mathbf{A}'\Psi^{-2}\mathbf{A} + \mathbf{I})\mathbf{A}'$$

eli kertomalla matriisilla $\mathbf{R}^{-1}$ oikealta saadaan

$$\mathbf{A}'\Psi^{-2} = (\mathbf{A}'\Psi^{-2}\mathbf{A} + \mathbf{I})\mathbf{A}'\mathbf{R}^{-1} = (\mathbf{A}'\Psi^{-2}\mathbf{A} + \mathbf{I})\mathbf{K} ,$$

josta **K** ratkeaa muodossa

$$\mathbf{K} = (\mathbf{A}'\Psi^{-2}\mathbf{A} + \mathbf{I})^{-1}\mathbf{A}'\Psi^{-2} .$$

Faktoripistemäärien kerroinmatriisi **K** saadaan siis lasketuksi kääntämällä "suuren" $p \times p$-korrelaatiomatriisin **R** asemasta "pieni" $r \times r$-matriisi.

Tulokset on johdettu olettaen faktorit ortogonaalisiksi. Jos faktorit korreloivat eli $\mathrm{E}(\mathbf{FF}')=\Phi$, saadaan vastaavalla tavalla $\mathbf{K}=\Phi\mathbf{A}'\mathbf{R}^{-1}$ ja Ledermannin lyhennetyssä muodossa

$$\mathbf{K} = (\mathbf{A}'\Psi^{-2}\mathbf{A} + \Phi^{-1})^{-1}\mathbf{A}'\Psi^{-2} .$$

Survossa faktoripistemäärien kerroinmatriisi **K** lasketaan sukrolla `/FCOEFF` ortogonaalisten faktorien tapauksessa ja sukrolla `/FTCOEFF` vinorotaation jälkeen.

5.5.1 Esimerkki

Jatkamme rotaatiomenetelmien yhteydessä luodun 300 havainnon XYZ-otoksen käsittelyä laskemalla kosinirotaation pohjalta faktoripistemäärät. Rotaatiossa saatiin seuraavat tulokset:

```
 53  1 SURVO 84C EDITOR Fri Mar 25 09:54:16 1994          D:\M\MEN\ 240 100 0
144 *ROTATE FACT.M,3,CUR+1 / ROTATION=COS     RESULTS=100_
145 *Rotated factor matrix AFACT.M=FACT.M*inv(TFACT.M)'
146 *              F1     F2     F3 Sumsqr
147 *X1         0.816  0.000 -0.000  0.665
148 *X2         0.620  0.393 -0.237  0.595
149 *X3        -0.601  0.570 -0.076  0.692
150 *X4         0.583  0.159  0.081  0.371
151 *Y1        -0.000  0.897 -0.000  0.805
152 *Y2         0.042  0.737 -0.232  0.598
153 *Y3         0.335 -0.449  0.350  0.436
154 *Y4         0.197  0.321 -0.140  0.161
155 *Z1         0.000 -0.000  0.651  0.424
156 *Z2        -0.130 -0.187  0.587  0.397
157 *Z3         0.323 -0.069 -0.514  0.373
158 *Z4        -0.259  0.222  0.256  0.182
159 *Sumsqr     2.091  2.247  1.363  5.700
160 *
161 *Rotation matrix TFACT.M
162 *              F1     F2     F3
163 *F1         0.666  0.933 -0.182
164 *F2         0.741 -0.234  0.576
165 *F3        -0.091  0.273  0.797
166 *
167 *Factor correlation matrix RFACT.M
168 *              F1     F2     F3
169 *F1         1.000  0.423  0.233
170 *F2         0.423  1.000 -0.087
171 *F3         0.233 -0.087  1.000
172 *
173 *The factor structure matrix SFACT.M is obtained by the commands:
174 *MAT SFACT.M=AFACT.M*RFACT.M  / *SFACT.M~A*RFACT 12*3
175 *MAT LOAD SFACT.M,12.123,CUR+1
176 *
```

Nähdäksemme, kuinka tarkasti faktorien arvot pystytään regressiomenetelmällä saamaan, tulemme lopuksi laskemaan alkuperäisten muuttujien ja faktorien väliset korrelaatiokertoimet. Näiden korrelaatiokertoimien estimaatit saadaan suoraan rotaation yhteydessä matriisina $\mathbf{S}=\mathbf{A}\Phi$, joka voidaan laskea ja ottaa esille suoraan ROTATE-operaation tulosten loppuun kirjoittamilla matriisikäskyillä (rivit 174-175 edellä). Teemme tämän ennen faktoripistemäärien laskemista.

```
  1  1 SURVO 84C EDITOR Fri Mar 25 10:03:21 1994          D:\M\MEN\ 240 100 0
173 *The factor structure matrix SFACT.M is obtained by the commands:
174 *MAT SFACT.M=AFACT.M*RFACT.M  / *SFACT.M~A*RFACT 12*3
175 *MAT LOAD SFACT.M,12.123,CUR+1
176 *MATRIX SFACT.M
177 *A*RFACT
178 *///          F1     F2     F3
179 *X1         0.816  0.345  0.190
180 *X2         0.730  0.675 -0.127
181 *X3        -0.377  0.323 -0.266
182 *X4         0.669  0.398  0.203
183 *Y1         0.379  0.897 -0.078
184 *Y2         0.300  0.775 -0.286
185 *Y3         0.227 -0.338  0.467
186 *Y4         0.300  0.416 -0.122
187 *Z1         0.152 -0.057  0.651
188 *Z2        -0.072 -0.294  0.573
189 *Z3         0.174  0.113 -0.433
190 *Z4        -0.106  0.091  0.176
191 *
```

Faktoripistemäärien kerroinmatriisi **K** lasketaan rivin 192 /FTCOEFF-sukrokomennolla, jossa lähtötietoina annetaan rotatoitu faktorimatriisi **A** (AFACT.M), rotaatiomatriisi **T** (TFACT.M) ja alkuperäisten muuttujien keskiarvojen ja hajontojen matriisi MSN.M . Viimeinen parametri (FCOEFF.M) ilmoittaa tulosta **K** vastaavan matriisitiedoston nimen. Itse asiassa /FTCOEFF ilman mitään parametreja tekee saman, sillä tässä käytetyt parametrit ovat kaikki oletusarvoja.

Ainoana näkyvänä tietona /FTCOEFF kirjoittaa (riville 193) ohjeen, miten laskea saadun kerroinmatriisin avulla faktoripistemäärät. Tarvittava LINCO-operaatio on käynnistetty rivillä 195. Lopuksi on laskettu kaikkien muuttujien (12 alkuperäistä ja 3 faktoripistemäärää) korrelaatiomatriisi ja poimittu näkyville alkuperäisten muuttujien ja juuri laskettujen faktoripistemäärien korrelaatiokertoimet:

```
 38  1 SURVO 84C EDITOR Fri Mar 25 10:05:34 1994          D:\M\MEN\ 240 100 0
191 *
192 */FTCOEFF AFACT.M,TFACT.M,MSN.M,FCOEFF.M
193 *Use FCOEFF.M for factor scores by LINCO <data>,FCOEFF.M(F1,F2,...)
194 *
195 *LINCO XYZ,FCOEFF.M(F1,F2,F3)
196 *CORR XYZ
197 *MAT LOAD CORR.M(*,13:15),##.###,CUR+1
198 *MATRIX CORR.M
199 *R(XYZ)
200 *///          F1     F2     F3
201 *X1         0.890  0.367  0.227
202 *X2         0.751  0.698 -0.139
203 *X3        -0.492  0.388 -0.348
204 *X4         0.709  0.418  0.234
205 *Y1         0.411  0.957 -0.095
206 *Y2         0.325  0.814 -0.335
207 *Y3         0.245 -0.362  0.569
208 *Y4         0.328  0.440 -0.142
209 *Z1         0.168 -0.063  0.783
210 *Z2        -0.080 -0.309  0.682
211 *Z3         0.192  0.118 -0.547
212 *Z4        -0.117  0.099  0.221
213 *F1         1.000  0.438  0.255
214 *F2         0.438  1.000 -0.161
215 *F3         0.255 -0.161  1.000
216 *
```

Kolme viimeistä riviä vastaa faktorien korrelaatiomatriisia Φ, jonka estimaatti rotaation jälkeen RFACT.M on riveillä 167-171. Samalla tavalla korrelaatioi-

den riveillä 201-212 tulisi vastata matriisia **S** (`SFACT.M`) riveillä 179-190. Vastaavuus ei ole täydellistä, mutta se osoittaa, että regressiomenetelmä toimii tässä tapauksessa varsin tyydyttävästi.

Vertailun vuoksi voidaan vielä ottaa esille simulointiesimerkin lähtökohtana olleet **A**- ja Φ-matriisit ja laskea niiden avulla todellinen **S**:

```
 24  1 SURVO 84C EDITOR Fri Mar 25 10:36:34 1994          D:\M\MEN\ 240 100 0
216 *
217 *MAT S=A*PHI        / *S~A*PHI 12*3
218 *MAT LOAD S,##.###,CUR+1_
219 *MATRIX S
220 *A*PHI
221 *///          F1     F2     F3
222 *X1        0.800  0.320  0.160
223 *X2        0.780  0.600 -0.090
224 *X3       -0.400  0.260 -0.170
225 *X4        0.600  0.390  0.180
226 *Y1        0.360  0.900 -0.090
227 *Y2        0.340  0.760 -0.250
228 *Y3        0.160 -0.410  0.410
229 *Y4        0.320  0.380  0.010
230 *Z1        0.120 -0.060  0.600
231 *Z2       -0.180 -0.330  0.480
232 *Z3        0.220  0.160 -0.340
233 *Z4       -0.120  0.150  0.210
234 *
```

5.6 Transformaatioanalyysi

Nimityksen transformaatioanalyysi otti käyttöön *Yrjö Ahmavaara* vuonna 1954 väitöskirjassa julkaisemastaan faktorianalyysitulosten vertailumenetelmästä. Transformaatioanalyysia voi pitää ns. konfirmatorisen faktorianalyysin eräänä muotona. Transformaatioanalyysin yhteydessä tutkitaan paitsi vertailtavien faktorirakenteiden samankaltaisuutta myös erityisesti mahdollisia rakenne-eroja, jotka ilmenevät "poikkeavana transformoitumisena".

Transformaatioanalyysi vastaa myös osittain ns. *Prokrustes*-menetelmää (kts. esim. Seber ss. 253-6). Prokrustes oli kreikkalaisessa mytologiassa rosvo, joka sijoitti vieraat vuoteeseensa. Jos uhrit olivat liian lyhyitä, hän venytti heitä, jos taas liian pitkiä, hän pätki heidät vuoteen mittaisiksi. Prokrustes-menetelmässä ei kuitenkaan transformaatioanalyysin tyyliin kiinnitetä huomiota "uhrien kärsimyksiin" eli poikkeavaan transformoitumiseen.

Transformaatioanalyysin lähtökohtana on samoilla muuttujilla tehdyt faktorianalyysit kahdessa (tai joskus useammassa) ryhmässä, joista on laskettu $p\times r$-faktorimatriisit $\mathbf{A}_1$ ja $\mathbf{A}_2$. Transformaatioanalyysissa on kiinnostuksen kohteena faktorirakenteen invarianssi. Tämä invarianssi ei edellytä, että $\mathbf{A}_1=\mathbf{A}_2$ edes likimäärin, koska faktorimatriisit rotaatiomahdollisuudesta johtuen eivät ole yksikäsitteisiä. Kuitenkin invarianssi takaa, että olisi olemassa säännöllinen *$r\times r$-transformaatiomatriisi* $\mathbf{L}_{12}$, jolle on voimassa

$$\mathbf{A}_1\mathbf{L}_{12} \approx \mathbf{A}_2 \, .$$

$\mathbf{L}_{12}$ määrätään pienimmän neliösumman periaatteella vaatimalla, että *residuaalimatriisin*

$$\mathbf{E} = \mathbf{E}_{12} = \mathbf{A}_1\mathbf{L}_{12} - \mathbf{A}_2$$

alkioitten neliösumma $\|\mathbf{E}\|^2=\mathrm{tr}(\mathbf{EE}')$ eli *kokonaisresiduaali* on mahdollisimman pieni. Transformaatiomatriisille $\mathbf{L}_{12}$ saatetaan asettaa tässä minimoinnissa lisärajoituksia.

Kokonaisresiduaalin ohella transformaatioanalyysissa tarkastellaan myös residuaalikovarianssimatriisia $\mathbf{EE}'$ ja $\mathbf{E}$:n vaakariveittäin laskettuja neliösummia $\mathrm{diag}(\mathbf{EE}')$ eli muuttujakohtaisia residuaaleja. On syytä tähdentää, että residuaaleista näkyvät poikkeamat kuvaavat faktorirakenteiden eroja eikä esim. keskiarvoeroja, joita tutkitaan esim. T^2-testillä tai erotteluanalyysilla. Rakenne-eroista on käytetty nimitystä "poikkeava transformoituminen".

5.6.1 Ahmavaaran ratkaisu

Ahmavaaran alkuperäisessä esityksessä transformaatiomatriisille $\mathbf{L}_{12}$ ei aseteta mitään rajoituksia. Kyseessä on itse asiassa lineaarinen regressio-ongelma ja ratkaisu saadaan esim. muodossa

$$\mathbf{L}_{12} = (\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{A}_1'\mathbf{A}_2 .$$

Todistaaksemme tämän siirrytään yksinkertaisempiin merkintöihin $\mathbf{A}_1=\mathbf{A}$, $\mathbf{A}_2=\mathbf{B}$ ja $\mathbf{L}_{12}=\mathbf{L}$, jolloin on minimoitava

$$\mathrm{tr}(\mathbf{AL}-\mathbf{B})(\mathbf{AL}-\mathbf{B})'$$

matriisin **L** suhteen.

Otamme käyttöön matriisin **A** Gram-Schmidt-hajotelman

$$\mathbf{A} = \mathbf{QR} ,$$

missä **Q** on pystyriveittäin ortogonaalinen $p\times r$-matriisi eli

$$\mathbf{Q}'\mathbf{Q}=\mathbf{I}$$

ja **R** on $r\times r$-yläkolmiomatriisi. Olkoon edelleen

$$\mathbf{K} = \mathbf{RL} ,$$

jolloin $\mathbf{AL}=\mathbf{QRL}=\mathbf{QK}$ ja (nyt **K**:n suhteen) minimoitavaa lauseketta voidaan kehitellä seuraavasti:

$$\mathrm{tr}(\mathbf{AL}-\mathbf{B})(\mathbf{AL}-\mathbf{B})' = \mathrm{tr}(\mathbf{QK}-\mathbf{B})(\mathbf{QK}-\mathbf{B})'$$

$$= \mathrm{tr}(\mathbf{QKK}'\mathbf{Q}' - \mathbf{QKB}' - \mathbf{BK}'\mathbf{Q}' + \mathbf{BB}')$$

$$= \mathrm{tr}(\mathbf{KK}' - \mathbf{KB}'\mathbf{Q} - \mathbf{Q}'\mathbf{BK}' + \mathbf{BB}')$$

$$= \mathrm{tr}((\mathbf{K}-\mathbf{Q}'\mathbf{B})(\mathbf{K}-\mathbf{Q}'\mathbf{B})' - \mathbf{Q}'\mathbf{BB}'\mathbf{Q} + \mathbf{BB}') .$$

Lausekkeen viimeisestä esitysmuodosta näemme, että se saavuttaa pienimmän mahdollisen arvon, kun

$$\mathbf{K} = \mathbf{Q}'\mathbf{B} \quad \text{eli} \quad \mathbf{RL} = \mathbf{Q}'\mathbf{B} .$$

Tästä on periaatteessa yksinkertaisinta ratkaista **L** suoraan, koska **R** on kolmiomatriisi, muodossa

$$\mathbf{L} = \mathbf{R}^{-1}\mathbf{Q}'\mathbf{B} ,$$

mutta kirjallisuudessa tavallisempi esitystapa saadaan kertomalla yhtälö $\mathbf{RL}=\mathbf{Q}'\mathbf{B}$ vasemmalta matriisilla $\mathbf{R}'$

$$\mathbf{R}'\mathbf{RL} = \mathbf{R}'\mathbf{Q}'\mathbf{B}$$

ja kun muistetaan, että $\mathbf{A}=\mathbf{QR}$ sekä todetaan, että $\mathbf{A}'\mathbf{A}=\mathbf{R}'\mathbf{Q}'\mathbf{QR}=\mathbf{R}'\mathbf{R}$, niin päädytään yhtälöön

$$\mathbf{A}'\mathbf{AL} = \mathbf{A}'\mathbf{B} ,$$

josta **L** ratkeaa muodossa

$$\mathbf{L} = \mathbf{L}_{12} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{B} = (\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{A}_1'\mathbf{A}_2 .$$

Survossa ratkaisu saadaan aikaan suoraan matriisitulkin komennolla

```
MAT SOLVE L12 FROM A1*L12=A2 ,
```

joka silloin, kun yhtälöitä on enemmän kuin tuntemattomia, määrää ratkaisun pienimmän neliösumman keinolla. Numeerinen menettely vastaa tällöin edellä

esitettyä suoraan ortogonalisoidusta $\mathbf{A}_1$-matriisista johdettua. Käytännössä tehtävä suoritetaan sukrolla

`/TRAN-LEASTSQR A1,A2` ,

joka antaa transformaatiomatriisin $\mathbf{L}_{12}$ lisäksi residuaalimatriisin $\mathbf{E}_{12}$. `/TRAN` on transformaatioanalyysin keinoja sisältävä sukroperhe, josta saa lisätietoja aktivoimalla komennon `/TRAN-README` .

Ahmavaaran alkuperäisen ratkaisun heikkoutena voidaan pitää sitä, että tulos riippuu suunnasta, jossa transformaatio tehdään. On luonnollista vaatia ratkaisulta, että $\mathbf{L}_{21}=\mathbf{L}_{12}^{-1}$ eli $\mathbf{L}_{12}\mathbf{L}_{21}=\mathbf{I}$, sillä ihannetapauksessa

$$\mathbf{A}_1=\mathbf{A}_2\mathbf{L}_{21} = \mathbf{A}_2\mathbf{L}_{12}^{-1} .$$

Ko. ratkaisu ei täytä tätä ehtoa eikä myöskään sitä, että edes kokonaisresiduaalin tr(**EE**') arvo olisi suunnasta riippumaton.

5.6.2 Symmetrinen transformaatioanalyysi

On osoitettu (Mustonen, 1966), että maksimaalinen symmetria transformaatioanalyysin tuloksissa saavutetaan, kun $\mathbf{L}_{12}$ on ortogonaalinen.

Jos käytämme jälleen väliaikaisesti yksinkertaisempia merkintöjä $\mathbf{A}_1=\mathbf{A}$, $\mathbf{A}_2=\mathbf{B}$ ja $\mathbf{L}_{12}=\mathbf{L}$, tehtävänä on minimoida

$$f(\mathbf{L}) = \mathrm{tr}(\mathbf{AL}-\mathbf{B})(\mathbf{AL}-\mathbf{B})' \quad \text{ehdolla } \mathbf{L}'\mathbf{L}=\mathbf{I} .$$

Väitämme, että minimiarvo saavutetaan, kun

$$\mathbf{L} = \mathbf{UV}',$$

missä **U** ja **V** saadaan suoraan $r\times r$-matriisin $\mathbf{A}'\mathbf{B}$ singulaariarvohajotelmasta

$$\mathbf{A}'\mathbf{B} = \mathbf{UDV}'.$$

Osoittaaksemme tämän, toteamme, että

$$\begin{aligned} f(\mathbf{L}) &= \mathrm{tr}(\mathbf{AA}' + \mathbf{BB}') - \mathrm{tr}(\mathbf{ALB}') - \mathrm{tr}(\mathbf{BL}'\mathbf{A}') \\ &= \mathrm{tr}(\mathbf{AA}' + \mathbf{BB}') - \mathrm{tr}(\mathbf{L}(\mathbf{A}'\mathbf{B})') - \mathrm{tr}(\mathbf{L}'(\mathbf{A}'\mathbf{B})) \\ &= \mathrm{tr}(\mathbf{AA}' + \mathbf{BB}') - \mathrm{tr}(\mathbf{LVDU}') - \mathrm{tr}(\mathbf{L}'\mathbf{UDV}') \\ &= \mathrm{tr}(\mathbf{AA}' + \mathbf{BB}') - \mathrm{tr}(\mathbf{DU}'\mathbf{LV}) - \mathrm{tr}(\mathbf{DV}'\mathbf{L}'\mathbf{U}) \\ &\geq \mathrm{tr}(\mathbf{AA}' + \mathbf{BB}') - 2\,\mathrm{tr}(\mathbf{D}) , \end{aligned}$$

sillä sekä tr(**DU**'**LV**) että tr(**DV**'**L**'**U**) ovat muotoa tr(**DS**), missä **S** on ortogonaalinen matriisi ja

$\mathrm{tr}(\mathbf{DS}) \leq \mathrm{tr}(\mathbf{D})$.

Viimeinen epäyhtälö seuraa siitä, että

$\mathrm{tr}(\mathbf{D}) - \mathrm{tr}(\mathbf{DS}) = \mathrm{tr}(\mathbf{D} - \mathbf{DS}) = \mathrm{tr}(\mathbf{D}(\mathbf{I} - \mathbf{S})) \geq 0$,

koska ortogonaalisen matriisin **S** kaikki alkiot ovat ≤ 1 .

On helppo nähdä, että edellä saatu lausekkeen $f(\mathbf{L})$ alaraja saavutetaan vain, kun $\mathbf{U}'\mathbf{LV}=\mathbf{I}$ eli $\mathbf{L}=\mathbf{UV}'$.

Symmetrisessä transformaatioanalyysissa siis $\mathbf{L}_{12}$ saadaan yksinkertaisesti singulaariarvohajotelmasta $\mathbf{A}'_1\mathbf{A}_2=\mathbf{UDV}'$ muodossa $\mathbf{L}_{12}=\mathbf{UV}'$. Tällöin on selvää, että käänteistransformaation välittää $\mathbf{L}_{21}=\mathbf{VU}'$, jolloin symmetriaehto $\mathbf{L}_{21}\mathbf{L}_{12}=\mathbf{I}$ toteutuu.

Symmetrisessä mallissa jopa residuaalikovarianssimatriisit ovat samat suunnasta riippumatta eli

$$\mathbf{E}_{12}\mathbf{E}'_{12} = \mathbf{E}_{21}\mathbf{E}'_{21} ,$$

sillä

$$\begin{aligned}\mathbf{E}_{21}\mathbf{E}'_{21} &= (\mathbf{A}_2\mathbf{L}_{21} - \mathbf{A}_1)(\mathbf{L}'_{21}\mathbf{A}'_2 - \mathbf{A}'_1)\\ &= (\mathbf{A}_2 - \mathbf{A}_1\mathbf{L}_{12})(\mathbf{L}_{21}\mathbf{L}'_{21})(\mathbf{A}'_2 - \mathbf{L}'_{12}\mathbf{A}'_1)\\ &= (\mathbf{A}_1\mathbf{L}_{12} - \mathbf{A}_2)(\mathbf{A}_1\mathbf{L}_{12} - \mathbf{A}_2)'\\ &= \mathbf{E}_{12}\mathbf{E}'_{12} .\end{aligned}$$

Siten myös muuttujakohtaiset residuaalit diag(**EE**') ja kokonaisresiduaali tr(**EE**') ovat transformaation suunnasta riippumatta samat.

Survossa symmetrinen transformaatioanalyysi tapahtuu sukrokomennolla
`/TRAN-SYMMETR A1,A2` .

5.6.3 Esimerkki 1

Palataan 3 muuttujan ja 300 havainnon otokseen `XYZ`, jota käsiteltiin viimeksi rotaation yhteydessä ja jossa todettiin kosinirotaation antaman tuloksen silmämääräisesti vastaavan "todellista" faktorimatriisia:

```
 37  1 SURVO 84C EDITOR Sun Apr 03 15:33:54 1994          D:\M\MEN\ 100 100 0
 39 *
 40 */MNSIMUL R,*,XYZ,300  / RND=rand(23031994)
 41 *CORR XYZ
 42 *FACTA CORR.M,3
 43 *ROTATE FACT.M,3,CUR+1 / ROTATION=COS_
 44 *Rotated factor matrix AFACT.M=FACT.M*inv(TFACT.M)'
 45 *             F1     F2     F3 Sumsqr
 46 *X1         0.816  0.000 -0.000  0.665
 47 *X2         0.620  0.393 -0.237  0.595
 48 *X3        -0.601  0.570 -0.076  0.692
 49 *X4         0.583  0.159  0.081  0.371
 50 *Y1        -0.000  0.897 -0.000  0.805
 51 *Y2         0.042  0.737 -0.232  0.598
 52 *Y3         0.335 -0.449  0.350  0.436
 53 *Y4         0.197  0.321 -0.140  0.161
 54 *Z1         0.000 -0.000  0.651  0.424
 55 *Z2        -0.130 -0.187  0.587  0.397
 56 *Z3         0.323 -0.069 -0.514  0.373
 57 *Z4        -0.259  0.222  0.256  0.182
 58 *Sumsqr     2.091  2.247  1.363  5.700
 59 *
```

Tarkistamme saadun rotaatioratkaisun (`AFACT.M`) ja alkuperäisen faktorirakenteen (`A`) vastaavuuden sukrolla `/TRAN-LEASTSQR`, joka laskee Ahmavaaran alkuperäisen transformaatioanalyysin mukaiset tulokset:

```
  1  1 SURVO 84C EDITOR Sun Apr 03 15:48:37 1994          D:\M\MEN\ 100 100 0
 62 *
 63 */TRAN-LEASTSQR AFACT.M,A
 64 *MAT LOAD L.M,##.###,END+2  / Transformation matrix
 65 *MAT LOAD E.M,##.###,END+2  / Residual matrix
 66 *
 67 *MATRIX L.M
 68 *Transformation_matrix
 69 *///          F1     F2     F3
 70 *F1        0.989 -0.008  0.004
 71 *F2        0.023  0.968  0.049
 72 *F3       -0.061 -0.013  0.880
 73 *
 74 *MATRIX E.M
 75 *Residual_matrix
 76 *///          F1     F2     F3
 77 *X1        0.007 -0.007  0.004
 78 *X2       -0.064  0.078  0.013
 79 *X3        0.023  0.058 -0.042
 80 *X4        0.075 -0.052 -0.018
 81 *Y1        0.020 -0.031  0.044
 82 *Y2       -0.027  0.016  0.032
 83 *Y3       -0.000  0.058 -0.012
 84 *Y4        0.010  0.011 -0.107
 85 *Z1       -0.040 -0.008 -0.027
 86 *Z2        0.031  0.012  0.007
 87 *Z3        0.049 -0.063 -0.054
 88 *Z4        0.033 -0.086 -0.065
 89 *
```

Sukro /TRAN-LEASTSQR tallettaa tulokset matriisitiedostoihin L.M ja E.M ja kirjoittaa valmiit MAT LOAD-komennot (tässä riveille 64-65), joiden aktivointi tuottaa rivit 67-88.

Näemme, että transformaatiomatriisi on lähellä yksikkömatriisia ja residuaalimatriisin alkiot ovat tyydyttävän lähellä nollaa. Residuaalien tilastolliseen tarkasteluun otamme kantaa esimerkissä 3.

5.6.4 Esimerkki 2

Tutkimme nyt kaavamaisen esimerkin kautta, miten "poikkeava transformoituminen" tulee esiin silloin, kun faktorirakenteissa on joidenkin muuttujien/faktorien suhteen selvät erot. Tarkastelemme kahta faktorimatriisia A1 ja A2, jotka eroavat toisistaan vain kahden alkion suhteen (etumerkki on muuttunut). Ko. alkiot näkyvät tehostettuina seuraavassa kaaviossa. A1 on tässä sama kuin A aikaisemmassa esimerkissä.

Voisimme ajatella, että aineistossa 1 muuttujien X3 ja Y3 luonne tai merkitys on ainakin osittain toinen kuin aineistossa 2.

```
 12  1 SURVO 84C EDITOR Sun Apr 03 16:02:49 1994          D:\M\MEN\ 100 100 0
  1 *
  2 *Faktorimatriisi 1:
  3 *MATRIX A1
  4 *///   F1    F2    F3
  5 *X1    0.8   0     0
  6 *X2    0.7   0.3  -0.2
  7 *X3   -0.6   0.5   0
  8 *X4    0.5   0.2   0.1
  9 *Y1    0     0.9   0
 10 *Y2    0.1   0.7  -0.2
 11 *Y3    0.3  -0.5   0.3
 12 *Y4    0.2   0.3   0
 13 *Z1    0     0     0.6
 14 *Z2   -0.2  -0.2   0.5
 15 *Z3    0.3   0    -0.4
 16 *Z4   -0.3   0.3   0.3
 17 *
 18 *Faktorimatriisi 2:
 19 *MATRIX A2
 20 *///   F1    F2    F3
 21 *X1    0.8   0     0
 22 *X2    0.7   0.3  -0.2
 23 *X3    0.6   0.5   0
 24 *X4    0.5   0.2   0.1
 25 *Y1    0     0.9   0
 26 *Y2    0.1   0.7  -0.2
 27 *Y3    0.3   0.5   0.3
 28 *Y4    0.2   0.3   0
 29 *Z1    0     0     0.6
 30 *Z2   -0.2  -0.2   0.5
 31 *Z3    0.3   0    -0.4
 32 *Z4   -0.3   0.3   0.3
 33 *
 34 *MAT SAVE A1
 35 *MAT SAVE A2
 36 *
```

Jos oletamme faktorirakenteet ortogonaalisiksi, faktorirakenteiden vertaamiseen on käytettävissä transformaatioanalyysin symmetrinen muoto. Käytämme sukroa /TRAN-SYMMETR, joka tuottaa seuraavat tulokset:

```
  1  1 SURVO 84C EDITOR Sun Apr 03 16:14:53 1994          D:\M\MEN\ 100 100 0
 36 *
 37 */TRAN-SYMMETR A1,A2
 38 *MAT LOAD L.M,##.###,END+2  / Transformation matrix
 39 *MAT LOAD E.M,##.###,END+2  / Residual matrix
 40 *
 41 *MATRIX L.M
 42 *Transformation_matrix
 43 *///          F1     F2     F3
 44 *F1        0.996 -0.086  0.014
 45 *F2        0.087  0.991 -0.101
 46 *F3       -0.005  0.102  0.995
 47 *
 48 *MATRIX E.M
 49 *Residual_matrix
 50 *///          F1     F2     F3
 51 *X1       -0.003 -0.069  0.011
 52 *X2        0.024 -0.083 -0.020
 53 *X3       -1.154  0.047 -0.059
 54 *X4        0.015 -0.035 -0.014
 55 *Y1        0.078 -0.008 -0.091
 56 *Y2        0.062 -0.035 -0.068
 57 *Y3       -0.046 -0.991  0.053
 58 *Y4        0.025 -0.020 -0.027
 59 *Z1       -0.003  0.061 -0.003
 60 *Z2       -0.019  0.070  0.015
 61 *Z3        0.001 -0.066  0.006
 62 *Z4        0.026  0.054 -0.036
 63 *
```

Transformaatiomatriisi pysyy miltei yksikkömatriisina ja residuaalimatriisissa muuttujien X3 ja Y3 poikkeva käyttäytyminen näkyy ilmiselvästi.

5.6.5 Esimerkki 3

Poikkeavaa transformoitumista on hankala tutkia tilastollisesti. Periaatteessa ongelmaa voi lähestyä konfirmatorisen faktorianalyysin suunnasta.

Tässä yhteydessä testaamme transformaatioanalyysin residuaaleja satunnaistamisperiaatteella, jolloin residuaalien jakaumat samanrakenteisten faktoriratkaisujen vertailussa arvioidaan aineistokohtaisesti. Tätä varten on tehty kaksi sukroa, /TRAN-LSTRES ja /TRAN-SYMTRES, jotka määräävät residuaalimatriisin alkioiden keskivirheet raa'alla Monte Carlo-menetelmällä, edellinen Ahmavaaran alkuperäisen mallin ja jälkimmäinen symmetrisen mallin yhteydessä. Ko. sukrojen toiminnasta saa lisätietoja aktivoimalla ne ilman parametreja.

Alustavat kokeet ovat osoittaneet, että kunkin yksittäisen residuaalin jakauma on likimain normaalinen odotusarvolla 0, mutta niiden hajonnat vaihtelevat ei ainoastaan otoksien koosta vaan myös faktorirakenteesta riippuen.

Tarkastelemme tässä esimerkkinä sukroa /TRAN-SYMTRES. Oletamme, että $\mathbf{A}_1$ ja $\mathbf{A}_2$ ovat kaksi (ortogonaalista) faktoriratkaisua, joista edellinen on saatu N_1 havainnon otoksen perusteella ja jälkimmäinen joko N_2 havainnon otoksen perusteella tai $\mathbf{A}_2$ on annettu hypoteettinen faktorimatriisi. Oletamme edelleen, että on tehty symmetrinen transformaatioanalyysi, joka on antanut residuaalimatriisin $\mathbf{E}=\mathbf{A}_1\mathbf{L}-\mathbf{A}_2$.

Residuaalien jakaumaa simuloidaan olettaen, että faktoriratkaisut $\mathbf{A}_1$ ja $\mathbf{A}_2$ ovat samat (rotaatiota vaille). Ensin luodaan matriisin $\mathbf{A}_2$ mukaista faktorimallia noudattava N_1 havainnon otos ja tästä riippumaton saman mallin mukainen N_2 havainnon otos. Näille otoksille muodostetaan suurimman uskottavuuden faktoriratkaisut FACTA-operaatiolla ja jälkimmäinen (valinnanvaraisesti) rotatoidaan ROTATE-operaatioilla. Tuloksista lasketaan symmetrisen transformaatiomallin mukainen residuaalimatriisi. Jos $\mathbf{A}_2$ on hypoteettinen matriisi, jälkimmäistä otosta ei luoda, vaan käytetään matriisia $\mathbf{A}_2$ sellaisenaan.

Sukro toistaa tämän kokeen niin monta kertaa kuin halutaan ja tallettaa jokaisen kokeen tulokset ($p\times r$ residuaalia) valittuun Survon havaintotiedostoon yhtenä havaintovektorina. Kun sovittu määrä toistokokeita on tehty, sukro laskee residuaalien keskivirheet ja kokoaa ne faktorimatriisin muotoiseksi matriisiksi.

Koska koetoistoja tarvitaan yleensä ainakin 100, menettely on melko raskas, sillä vaatiihan se yhtä monta multinormaalisten otosten generointia sekä niiden faktori- ja transformaatioratkaisut. Sukro ilmoittaakin jokaisen koetoiston jälkeen arvion jäljellä olevasta laskenta-ajasta tunteina ja minuutteina.
Tutkija voi keskeyttää ajon koetoistojen välissä (napilla S), jolloin tulokset lasketaan siihen asti kertyneiden tietojen perusteella.

Menettelyn havainnollistamiseksi otamme lähtökohdaksi jälleen 12×3-faktori-

matriisin A:

```
 11  1 SURVO 84C EDITOR Mon Apr 04 10:54:28 1994          D:\M\MEN\ 200 100 0
  1 *
  2 *Faktorimatriisi:
  3 *MATRIX A
  4 *///   F1    F2    F3
  5 *X1    0.8   0     0
  6 *X2    0.7   0.3  -0.2
  7 *X3   -0.6   0.5   0
  8 *X4    0.5   0.2   0.1
  9 *Y1    0     0.9   0
 10 *Y2    0.1   0.7  -0.2
 11 *Y3    0.3  -0.5   0.3
 12 *Y4    0.2   0.3   0
 13 *Z1    0     0     0.6
 14 *Z2   -0.2  -0.2   0.5
 15 *Z3    0.3   0    -0.4
 16 *Z4   -0.3   0.3   0.3
 17 *
 18 *MAT SAVE A_
 19 *
```

Oletamme faktorit korreloimattomiksi ja laskemme faktorianalyysin perusyhtälön avulla tätä faktorimatriisia vastaavan korrelaatiomatriisin R (rivit 21-27). Tämän jälkeen luomme ko. korrelaatiorakennetta vastaavan 100 havainnon otoksen XYZ100 multinormaalijakaumasta, laskemme korrelaatiomatriisin (CORR.M) sekä teemme tämän pohjalta suurimman uskottavuuden faktoroinnin ja Varimax-rotaation (rivit 29-32):

```
 22  1 SURVO 84C EDITOR Mon Apr 04 11:00:02 1994          D:\M\MEN\ 200 100 0
 19 *
 20 *Lasketaan korrelaatiomatriisi R perusyhtälöstä R=A*A'+PSI :
 21 *MAT DIM A /* rowA=12 colA=3
 22 *MAT R=MMT(A)      / *R~A*A' S12*12
 23 *MAT D=VD(R)        / *D~VD(A*A') 12*1
 24 *MAT D=DV(D)        / *D~DV(VD(A*A')) D12*12
 25 *MAT I=IDN(rowA,rowA)
 26 *MAT PSI!=I-D       / *PSI~IDN-DV(VD(A*A')) D12*12
 27 *MAT R=R+PSI        / *R~A*A'+PSI S12*12
 28 *
 29 */MNSIMUL R,*,XYZ100,100 / RND=rand(41994)
 30 *CORR XYZ100
 31 *FACTA CORR.M,3
 32 *ROTATE FACT.M,3,CUR+1_
 33 *Rotated factor matrix AFACT.M=FACT.M*TFACT.M
 34 *             F1     F2     F3 Sumsqr
 35 *X1       -0.315  0.258  0.669  0.614
 36 *X2        0.044  0.475  0.637  0.633
 37 *X3        0.732 -0.136 -0.349  0.677
 38 *X4        0.019 -0.033  0.421  0.179
 39 *Y1        0.820 -0.103  0.165  0.711
 40 *Y2        0.655  0.414  0.302  0.692
 41 *Y3       -0.582 -0.186  0.290  0.458
 42 *Y4        0.217 -0.027  0.210  0.092
 43 *Z1       -0.058 -0.511 -0.035  0.266
 44 *Z2       -0.069 -0.783  0.030  0.620
 45 *Z3        0.018  0.629  0.073  0.401
 46 *Z4        0.249 -0.435 -0.132  0.269
 47 *Sumsqr    2.197  1.989  1.425  5.611
 48 *
```

Kopioimme rotatoidun faktorimatriisin AFACT.M matriisiksi AT ja vertaamme tätä alkuperäiseen faktorimatriisiin symmetrisen transformaatioanalyysin keinoin:

```
  1  1 SURVO 84C EDITOR Mon Apr 04 11:19:35 1994          D:\M\MEN\ 200 100 0
 51 *
 52 *MAT AT=AFACT.M    / *AT~A 12*3
 53 */TRAN-SYMMETR AT,A
 54 *MAT LOAD L.M,##.###,END+2  / Transformation matrix
 55 *MAT LOAD E.M,##.###,END+2  / Residual matrix
 56 *
 57 *MATRIX L.M
 58 *Transformation_matrix
 59 *///          F1     F2     F3
 60 *F1       -0.304  0.950 -0.067
 61 *F2        0.319  0.035 -0.947
 62 *F3        0.898  0.309  0.314
 63 *
 64 *MATRIX E.M
 65 *Residual_matrix
 66 *///          F1     F2     F3
 67 *X1       -0.021 -0.083 -0.013
 68 *X2        0.010 -0.044 -0.052
 69 *X3        0.021  0.083 -0.030
 70 *X4       -0.138 -0.053  0.063
 71 *Y1       -0.134 -0.073  0.095
 72 *Y2        0.104  0.031 -0.141
 73 *Y3        0.078  0.030  0.006
 74 *Y4       -0.086 -0.030  0.077
 75 *Z1       -0.177 -0.084 -0.123
 76 *Z2       -0.002  0.116  0.256
 77 *Z3       -0.039  0.062 -0.174
 78 *Z4       -0.033 -0.120  0.054
 79 *
```

Saadaksemme käsityksen residuaalien keskivirheistä teemme 100-kertaisen simulointikokeen sukrolla /TRAN-SYMTRES:

```
  1  1 SURVO 84C EDITOR Mon Apr 04 11:24:48 1994          D:\M\MEN\ 200 100 0
 80 *.....................................................................
 81 */TRAN-SYMTRES A,-,100,*,ARES,100,1200001
 82 *Simulated residuals in Survo data file ARES.SVO
 83 *MAT LOAD ARES,##.###,END+2 / Standard errors of residuals
 84 *
 85 *MATRIX ARES
 86 *Standard_errors_of_residuals_(N=100)
 87 *///          F1     F2     F3
 88 *X1        0.056  0.067  0.088
 89 *X2        0.062  0.068  0.080
 90 *X3        0.066  0.069  0.076
 91 *X4        0.081  0.092  0.098
 92 *Y1        0.065  0.053  0.073
 93 *Y2        0.072  0.065  0.093
 94 *Y3        0.094  0.081  0.104
 95 *Y4        0.084  0.099  0.141
 96 *Z1        0.098  0.090  0.152
 97 *Z2        0.095  0.088  0.136
 98 *Z3        0.088  0.097  0.140
 99 *Z4        0.073  0.088  0.114
100 *
```

Kun vertaamme residuaaleja (rivit 67-78) niiden arvioituihin keskivirheisiin (rivit 88-99), voimme todeta, ettei mitään poikkeavaa transformoitumista esiinny, kuten sopi odottaakin. Suurimmat residuaalin ja keskivirheen suhteet ovat suuruusluokkaa 2 ja niitäkin on vain muutamia.

Simulointi käynnistettiin rivillä 81 olevalla komennolla

```
/TRAN-SYMTRES A,-,100,*,ARES,100,1200001 ,
```

jonka parametrit tulkitaan seuraavasti:

A	Jakaumaa simuloidaan faktorimatriisin A mukaisesti.
-	Rotaatiota ei tehdä (koska verrataan teoreettiseen matriisiin A).

`100`	Otoskoko N_1 on 100.
`*`	Toinen otoskoko on "ääretön" (eli vertailu teoreettiseen `A`).
`ARES`	Tulosmatriisin nimi
`100`	Simulointikertojen lukumäärä on 100.
`1200001`	Ensimmäinen simulointi tehdään generaattorilla `rand(1200001)`.

Verrattaessa kahta faktorimatriisia $\mathbf{A}_1$ ja $\mathbf{A}_2$, jotka on laskettu otoskoilla N_1 ja N_2, residuaalimatriisin $\mathbf{E}=\mathbf{A}_1\mathbf{L}-\mathbf{A}_2$ simulointi tapahtuu muotoa

`/TRAN-SYMTRES` $\mathbf{A}_2$,<rotaatio>,N_1,N_2,<tulostiedosto>,N,<rand>

olevalla komennolla. Tässä parametrilla <rotaatio> ovat vain vaihtoehdot `VARIMAX` ja `-` .

Palataksemme vielä äskeiseen esimerkkiin, simuloidut residuaalit on talletettu muuttujina `E1`,`E2`,...,`E36` (Huom. $p\times r$=36 tässä tapauksessa) Survon havaintotiedostoon `ARES.SVO` . Kunkin havainnon alussa on `CASE`-niminen muuttuja, jossa on käytetyn satunnaislukugeneraattorin indeksi. Siis tässä tapauksessa havainnot on nimetty 1200001, 1200002,...

Erillisten residuaalien ohella on syytä tarkastella muuttujakohtaisia residuaaleja (**E**:n vaakarivien neliösummat) ja kokonaisresiduaalia (kaikkien alkioiden neliösumma). Kun simulointikertojen määrä on riittävä, on tulostiedostosta (tässä `ARES.SVO`) mahdollista määrätä likimain ko. neliösummien ylärajat valituilla kriittisillä tasoilla. Esimerkissämme voimme laskea neliösummat ensin "havaitussa tilanteessa" eli matriisille `E.M` matriisiketjulla `MATRUN SUM2`:

```
 23  1 SURVO 84C EDITOR Mon Apr 04 17:47:59 1994          D:\M\MEN\ 200 100 0
100 *
101 *MATRUN SUM2,E.M,##.###_
102 *
103 *MATRIX &G
104 *"E.M_with_sums_of_squares_by_rows_and_columns"
105 *///          F1     F2     F3 sumsqr
106 *X1       -0.021 -0.083 -0.013  0.008
107 *X2        0.010 -0.044 -0.052  0.005
108 *X3        0.021  0.083 -0.030  0.008
109 *X4       -0.138 -0.053  0.063  0.026
110 *Y1       -0.134 -0.073  0.095  0.032
111 *Y2        0.104  0.031 -0.141  0.032
112 *Y3        0.078  0.030  0.006  0.007
113 *Y4       -0.086 -0.030  0.077  0.014
114 *Z1       -0.177 -0.084 -0.123  0.053
115 *Z2       -0.002  0.116  0.256  0.079
116 *Z3       -0.039  0.062 -0.174  0.035
117 *Z4       -0.033 -0.120  0.054  0.018
118 *sumsqr    0.096  0.065  0.156  0.318
119 *
```

Simulointituloksista lasketaan tarvittavat suureet esim. `VARSTAT`- ja `STAT`-operaatioilla seuraavasti:

```
  44  1 SURVO 84C EDITOR Mon Apr 04 17:51:20 1994          D:\M\MEN\ 200 100 0
119 *
120 *MASK=-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
121 *VARSTAT ARES,Summa2,SUM,2
122 *STAT ARES,CUR+1 / VARS=Summa2 FRACTILES=0.9,0.95_
123 *Basic statistics: ARES N=100
124 *Variable: Summa2
125 *min=0.141451 in obs.#40 (1200040)
126 *max=0.845863 in obs.#35 (1200035)
127 *mean=0.303377 stddev=0.112814 skewness=1.751569 kurtosis=5.396984
128 *lower_Q=0.225    median=0.286765 upper_Q=0.369643
129 *fractile(0.9)=0.429545
130 *fractile(0.95)=0.5125
131 *up.limit    f     % class width=0.05
132 *    0.15    1   1.0 *
133 *    0.2    12  12.0 ************
134 *    0.25   25  25.0 *************************
135 *    0.3    17  17.0 *****************
136 *    0.35   15  15.0 ***************
137 *    0.4    14  14.0 **************
138 *    0.45   11  11.0 ***********
139 *    0.5     0   0.0
140 *    0.55    2   2.0 **
141 *    0.6     1   1.0 *
142 *    0.65    0   0.0
143 *    0.7     0   0.0
144 *    0.75    1   1.0 *
145 *    0.8     0   0.0
146 *    0.85    1   1.0 *
147 *
```

Rivillä 121 olevalla VARSTAT-komennolla lasketaan aktiivisten muuttujien neliösumma muuttujana Summa2. Aktiiviset muuttujat (E1-E36) on osoitettu edellisen rivin MASK-täsmennyksellä eli tuloksena saadaan transformaatioanalyysin kokonaisresiduaali kunkin simulointikokeen osalta.

Kokonaisresiduaalin jakaumasta syntyy käsitys rivin 122 STAT-komennolla, jossa lasketaan FRACTILES-täsmennystä käyttäen myös prosenttipisteet tasoilla 90% ja 95%.

Todetaan, että "havaittu" kokonaisresiduaali 0.318 (rivillä 118) on lähinnä mediaanin luokkaa eli mitään poikkeavaa transformoitumista ei tässä tapauksessa esiinny.

5.7 Faktorianalyysin kritiikistä

Faktorianalyysin suosio ja sen käyttö on vaihdellut melkoisesti viime vuosikymmenten aikana. Useat tilastotieteen tutkijat ovat suhtautuneet siihen hyvin kriittisesti. Arvostelu on ollut oikeutettua silloin, kun se on kohdistunut menetelmän käyttöön heppoisin aineistoin ja pohjatiedoin. On kuitenkin valitettavaa, että tarjotaan myös melko harhaanjohtavia väitteitä faktorianalyysin epämääräisyydestä ja kelvottomuudesta.

Esimerkkinä kohtuuttomasta kritiikistä käy se, mitä *G.A.Seber* esittää kirjassaan "Multivariate Observations" (1984). Hän käyttää tässä lähes 700 sivun teoksessaan itse faktorianalyysin selostamiseen noin 10 sivua, mutta faktorianalyysin kriittiseen tarkasteluun on uhrattu yli 10 sivua. Arvosteleva osuus perustuu voittopuolisesti simulointikokeisiin ja päätelmiin, joita *I.Francis* on tehnyt noin 10 vuotta aikaisemmin.

Francis tarkasteli kymmentä erilaista 10 muuttujaan ja 2-3 faktoriin perustuvaa faktorirakennetta, loi niiden perusteella toistuvasti keinotekoisia havaintoaineistoja ja laski näistä faktorianalyysin tulokset standardiohjelmilla.
Tutkittavan aineiston koko oli yleensä 50. Monet kokeista "epäonnistuivat" niin pahoin, että Seber päättelee (s.235): "In conclusion, it must be stated that if Factor Analysis is carried out, then the results must be interpreted with extreme caution. Even if the postulated model is true - and this is a very strong assumption - the chance of its recovery by present methods does not seem very great."

Ensimmäiseen lauseeseen voi tietenkin yhtyä; jokaisen tilastollisen menetelmän tuloksiin tulee suhtautua suurella varovaisuudella. Jälkimmäinen lause ei kuitenkaan pidä paikkaansa.

Eräs epäonnisista Francisin esimerkeistä oli viides (V), jossa kolmen faktorin matriisiksi **A** ja ominaisfaktorien keskihajonnoiksi valittiin

	A		diag(Ψ)
10	7	4	15
10	7	4	15
10	7	4	15
10	7	4	15
10	7	0	15
10	7	0	20
10	7	0	20
10	0	0	20
10	0	0	20
10	0	0	20

Huomattakoon, että Francisin kokeissa faktorimatriiseja ei skaalattu niin, että lataukset olisivat muuttujien ja faktorien korrelaatiokertoimia.

Faktoroinnissa Francis käytti mm. Jöreskogin silloista suurimman uskottavuuden ratkaisun antavaa ohjelmaa UFABY3 ja Seberin mukaan tässä tapauksessa ohjelma valitsi säännönmukaisesti faktoriluvuksi 1. Kun kuitenkin sovellettiin "oikeaa" lukumäärää 3, suuremmalla otoskoolla 250 saatiin ominaisfaktorien variansseille hyvät estimaatit, mutta estimoidut faktorimatriisit olivat kaikkea muuta kuin annettu **A** riippumatta käytetystä rotaatiomenetelmästä.

Jopa silloin kun lähtökohdaksi otettiin oikea kovarianssimatriisi (otoskoko ääretön), vain ominaisfaktorien varianssit saatiin oikein!

Tarkasteltakoon nyt aluksi tätä äärettömän otoskoon tilannetta mutta soveltaen aikaisemmin käyttämäämme skaalausta.

```
 14  1 SURVO 84C EDITOR Fri May 13 17:55:14 1994          D:\M\MEN\ 300 100 0
  1 *
  2 *MATRIX G
  3 *///   F1    F2    F3    C
  4 *X1    10     7     4   15
  5 *X2    10     7     4   15
  6 *X3    10     7     4   15
  7 *X4    10     7     4   15
  8 *X5    10     7     0   15
  9 *X6    10     7     0   20
 10 *X7    10     7     0   20
 11 *X8    10     0     0   20
 12 *X9    10     0     0   20
 13 *X10   10     0     0   20
 14 *
 15 *MAT SAVE G
 16 *MAT A!=G(*,1:3)    / A: matriisin G 3 ensimmäistä saraketta
 17 *MAT C!=G(*,4)_     / C: matriisin G 4. sarake
 18 *
```

Ensin on erotettu matriisitiedostosta `G` faktorimatriisi `A` ja `C` (ominaisfaktorien hajonnat). Näistä matriiseista lasketaan (muuttujien ja faktoreiden korrelaatiomatriisiksi) normeerattu faktorimatriisi `F` ja muuttujien korrelaatiomatriisi `R`:

```
  1  1 SURVO 84C EDITOR Fri May 13 17:57:10 1994          D:\M\MEN\ 300 100 0
 18 *
 19 *MAT TRANSFORM C BY X#*X#
 20 *MAT PSI2!=DV(C)    / ominaisvarianssit lävistäjämatriisina
 21 *MAT S=MMT(A)
 22 *MAT S=S+PSI2       / kovarianssimatriisi
 23 *MAT D=VD(S)
 24 *MAT TRANSFORM D BY sqrt(1/X#)
 25 *MAT D!=DV(D)       / muuttujien hajontojen käänteisarvot
 26 *MAT F=D*A          / F: uudelleen normeerattu faktorimatriisi
 27 *MAT R=D*S
 28 *MAT R!=R*D         / R: muuttujien korrelaatiomatriisi
 29 *_
```

Normeerattu faktorimatriisi `F` näyttää vaaka- ja pystyrivineliösummineen tällaiselta (tulostettuna matriisiketjulla `SUM2`):

```
 14  1 SURVO 84C EDITOR Fri May 13 18:04:23 1994          D:\M\MEN\ 300 100 0
 29 *
 30 *MATRUN SUM2 F_
 31 *
 32 *MATRIX &G
 33 *"F_with_sums_of_squares_by_rows_and_columns"
 34 *///             F1       F2       F3   sumsqr
 35 *X1         0.50637  0.35446  0.20255  0.42308
 36 *X2         0.50637  0.35446  0.20255  0.42308
 37 *X3         0.50637  0.35446  0.20255  0.42308
 38 *X4         0.50637  0.35446  0.20255  0.42308
 39 *X5         0.51709  0.36196  0.00000  0.39840
 40 *X6         0.42679  0.29875  0.00000  0.27140
 41 *X7         0.42679  0.29875  0.00000  0.27140
 42 *X8         0.44721  0.00000  0.00000  0.20000
 43 *X9         0.44721  0.00000  0.00000  0.20000
 44 *X10        0.44721  0.00000  0.00000  0.20000
 45 *sumsqr     2.25732  0.81209  0.16410  3.23351
 46 *
```

Näemme, että kolmannen faktorin osuus kokonaisvaihtelusta (10) on vaivaiset 1.6% ja yhteisvaihtelustakin (3.23) vain 5.6% . Lisäksi muuttujien kommunaliteetit ovat osittain hyvin alhaisia. Annettu rakenne ei siis ole järkevä.

On syytä myös kysyä, onko kyseessä lainkaan "yksinkertainen rakenne", jollaista faktorianalyysissa tavoitellaan. Katsomme, mitä tapahtuu, jos teemme tälle matriisille graafisen rotaation:

```
 40  1 SURVO 84C EDITOR Sat May 14 07:47:37 1994          D:\M\MEN\ 300 100 0
 46 *.....................................................................
 47 *ROTATE F,3,CUR+1  /  ROTATION=GRAPHICAL_
 48 *Rotated factor matrix AFACT.M=F*TFACT.M
 49 *             F1     F2     F3 Sumsqr
 50 *X1         0.618 -0.000  0.203  0.423
 51 *X2         0.618 -0.000  0.203  0.423
 52 *X3         0.618 -0.000  0.203  0.423
 53 *X4         0.618 -0.000  0.203  0.423
 54 *X5         0.631 -0.000  0.000  0.398
 55 *X6         0.521 -0.000  0.000  0.271
 56 *X7         0.521 -0.000  0.000  0.271
 57 *X8         0.366 -0.257  0.000  0.200
 58 *X9         0.366 -0.257  0.000  0.200
 59 *X10        0.366 -0.257  0.000  0.200
 60 *Sumsqr     2.872  0.197  0.164  3.234
 61 *
 62 *Rotation matrix saved as TFACT.M
 63 *Factors are orthogonal (RFACT.M=I).
 64 *
```

Rotaatiossa ei ole tapahtunut muuta kuin, että kahta ensimmäistä faktoria on kierretty 35 astetta, jolloin tarkkailemalla faktorien voimakkuuksia (neliösummat rivillä 60) havaitaan, että tosiasiassa faktoreita on vain yksi, joka selittää yhteisvaihtelusta 88%.

Soveltamalla graafisessa rotaatiossa Quartimax-kriteeriä johdonmukaisesti, päädytään edellistä tulosta muistuttavaan, jossa ensimmäisen faktorin selitysosuus on noussut 92 prosenttiin.

Niillä rippeillä, mitä kaksi muuta faktoria edustavat, voidaan vain leikitellä. Esim. pidettäessä teknisesti kiinni kolmesta faktorista, voi tietenkin todeta, että graafisella rotaatiolla saatu ratkaisu on rakenteeltaan yksinkertaisempi kuin Francisin alkuperäinen.

Ei kuitenkaan pidä ihmetellä, ettei esim. Varimax-kriteeri toimi kunnolla tässä tilanteessa, koska se "uskoo" jokaisen muuttujan merkittävyyteen ja

pidentää niitä vastaavat vektorit samanmittaisiksi:

```
 17  1 SURVO 84C EDITOR Sun May 15 12:46:00 1994          D:\M\MEN\ 300 100 0
 64 *.....................................................................
 65 *ROTATE F,3,CUR+1_    / ROTATION=VARIMAX
 66 *Rotated factor matrix AFACT.M=F*TFACT.M
 67 *             F1     F2     F3 Sumsqr
 68 *X1        0.278  0.586 -0.051  0.423
 69 *X2        0.278  0.586 -0.051  0.423
 70 *X3        0.278  0.586 -0.051  0.423
 71 *X4        0.278  0.586 -0.051  0.423
 72 *X5        0.305  0.501 -0.233  0.398
 73 *X6        0.252  0.413 -0.192  0.271
 74 *X7        0.252  0.413 -0.192  0.271
 75 *X8        0.400  0.192 -0.055  0.200
 76 *X9        0.400  0.192 -0.055  0.200
 77 *X10       0.400  0.192 -0.055  0.200
 78 *Sumsqr    1.009  2.077  0.147  3.234
 79 *
 80 *Rotation matrix saved as TFACT.M
 81 *Factors are orthogonal (RFACT.M=I).
```

Sen sijaan kosinirotaatio toimii "loistavasti", jos hyväksyy hyvin pientenkin kommunaliteettiarvojen muuttujat (tässä <0.20) ratkaisun kantavektoreiksi:

```
 17  1 SURVO 84C EDITOR Sun May 15 12:51:31 1994          D:\M\MEN\ 300 100 0
 82 *.....................................................................
 83 *ROTATE F,3,CUR+1_ / ROTATION=COS,0.19
 84 *Rotated factor matrix AFACT.M=F*inv(TFACT.M)'
 85 *             F1     F2     F3 Sumsqr
 86 *X1       -0.000  0.000  0.650  0.423
 87 *X2       -0.000  0.000  0.650  0.423
 88 *X3       -0.000  0.000  0.650  0.423
 89 *X4       -0.000  0.000  0.650  0.423
 90 *X5        0.631  0.000  0.000  0.398
 91 *X6        0.521 -0.000  0.000  0.271
 92 *X7        0.521 -0.000  0.000  0.271
 93 *X8        0.000  0.447  0.000  0.200
 94 *X9        0.000  0.447  0.000  0.200
 95 *X10       0.000  0.447  0.000  0.200
 96 *Sumsqr    0.941  0.600  1.692  3.234
 97 *
 98 *Rotation matrix saved as TFACT.M
 99 *Factor correlation matrix saved as RFACT.M
100 *
```

Rakenne on täysin puhdas; jokainen muuttuja latautuu vain yhdelle faktorille. Tämä tosin tapahtuu rankasti faktorien korreloivuuden kustannuksella:

```
 23  1 SURVO 84C EDITOR Sun May 15 12:54:14 1994          D:\M\MEN\ 300 100 0
100 *
101 *MAT LOAD RFACT.M,CUR+1_
102 *MATRIX RFACT.M
103 *RFACT
104 *///            F1       F2       F3
105 *F1       1.000000 0.819232 0.950279
106 *F2       0.819232 1.000000 0.778499
107 *F3       0.950279 0.778499 1.000000
108 *
```

Siis ilman mitään simulointikokeita on todettavissa, että Francisin valitsema faktorirakenne on käytännön kannalta mieletön. Kyseessä on tyypillinen yhden faktorin tapaus.

Jatkamme leikittelyä 3 faktorilla ja laskemme tämän mukaisen suurimman uskottavuuden ratkaisun "äärettömällä otoksella" eli suoraan Francisin rakenteen mukaisesta korrelaatiomatriisista:

```
  16  1 SURVO 84C EDITOR Sun May 15 13:57:31 1994          D:\M\MEN\ 300 100 0
108 *.....................................................................
109 *FACTA R,3,CUR+1_
110 *Factor analysis: Maximum Likelihood (ML) solution
111 *Factor matrix
112 *              F1     F2     F3    h^2
113 *X1         0.643 -0.080 -0.051  0.423
114 *X2         0.643 -0.080 -0.051  0.423
115 *X3         0.643 -0.080 -0.051  0.423
116 *X4         0.643 -0.080 -0.051  0.423
117 *X5         0.618  0.010  0.129  0.398
118 *X6         0.510  0.008  0.106  0.271
119 *X7         0.510  0.008  0.106  0.271
120 *X8         0.378  0.236 -0.037  0.200
121 *X9         0.378  0.236 -0.037  0.200
122 *X10        0.378  0.236 -0.037  0.200
123 *
```

Tulos ei todellakaan muistuta lähtökohtaa. Kuitenkin symmetrisellä transformaatioanalyysilla on helppo todeta rakenteiden täydellinen vastaavuus:

```
   1  1 SURVO 84C EDITOR Sun May 15 14:01:14 1994          D:\M\MEN\ 300 100 0
123 *
124 */TRAN-SYMMETR FACT.M,F
125 *MAT LOAD L.M,##.###,END+2  / Transformation matrix
126 *MAT LOAD E.M,##.########,END+2  / Residual matrix
127 *_
128 *MATRIX E.M
129 *Residual_matrix
130 *///               F1           F2           F3
131 *X1       -0.00000001  0.00000051 -0.00000086
132 *X2       -0.00000001  0.00000051 -0.00000086
133 *X3       -0.00000001  0.00000051 -0.00000086
134 *X4       -0.00000001  0.00000051 -0.00000086
135 *X5       -0.00000021 -0.00001079  0.00001854
136 *X6       -0.00000018  0.00000469 -0.00000850
137 *X7       -0.00000018  0.00000469 -0.00000850
138 *X8       -0.00000000  0.00000026 -0.00000044
139 *X9       -0.00000000  0.00000026 -0.00000044
140 *X10      -0.00000000  0.00000026 -0.00000044
141 *
```

Faktorianalyysi siis toimii "äärettömällä otoksella" juuri niin kuin pitääkin. Lievä kohina residuaaleissa osoittaa rakenteeseen kätkeytyvän likimaisen multikollineaarisuuden.

Francisin alkuperäinen matriisi on löytynyt täsmällisesti ortogonaalisella rotaatiolla. Millään standardirotaatiolla se ei voi kuitenkaan syntyä, koska Francisin rakenne ei ole "yksinkertainen".

Kahden kohtuuttoman pienen faktorin ansiosta eli multikollineaarisuudesta johtuen on turha odottaa, että oikea rakenne tulisi kovin tarkasti esiin pienillä otoksilla. Jos otoskoko on 250, tulokseksi saadaan esim.

```
  1  1 SURVO 84C EDITOR Sun May 15 14:16:29 1994          D:\M\MEN\ 300 100 0
141 *.....................................................................
142 */MNSIMUL R,*,FRANCIS5,250  / RND=rand(199401)
143 *CORR FRANCIS5
144 *FACTA CORR.M,3
145 */TRAN-SYMMETR FACT.M,F
146 *MAT LOAD L.M,##.###,END+2  / Transformation matrix
147 *MAT LOAD E.M,##.###,END+2  / Residual matrix
148 *_
149 *MATRIX E.M
150 *Residual_matrix
151 *///          F1     F2     F3
152 *X1       -0.060  0.028  0.122
153 *X2        0.034  0.005 -0.272
154 *X3       -0.022  0.049 -0.088
155 *X4       -0.032 -0.007  0.014
156 *X5        0.004 -0.049  0.386
157 *X6       -0.001  0.125 -0.228
158 *X7       -0.080 -0.031  0.077
159 *X8       -0.123  0.082 -0.025
160 *X9        0.438 -0.458 -0.031
161 *X10      -0.043  0.144 -0.029
162 *
```

Eräät residuaalit näyttävät suurilta, mutta tämä johtuu pelkästään siitä, että niiden keskivirheetkin ovat suuria:

```
 26  1 SURVO 84C EDITOR Mon May 16 08:17:44 1994          D:\M\MEN\ 300 100 0
162 *.....................................................................
163 */TRAN-SYMTRES F,-,250,*,FR5,100,155001
164 *Simulated residuals in Survo data file FR5.SVO
165 *MAT LOAD FR5,##.###,END+2_/ Standard errors of residuals
166 *
167 *MATRIX FR5
168 *Standard_errors_of_residuals_(N=100)
169 *///          F1     F2     F3
170 *X1        0.063  0.095  0.186
171 *X2        0.072  0.122  0.195
172 *X3        0.072  0.120  0.204
173 *X4        0.067  0.116  0.187
174 *X5        0.077  0.122  0.183
175 *X6        0.072  0.145  0.196
176 *X7        0.082  0.160  0.212
177 *X8        0.170  0.220  0.179
178 *X9        0.164  0.207  0.181
179 *X10       0.153  0.190  0.190
180 *
```

Kun otoskoko kasvaa, ratkaisu tarkentuu luonnollisesti kohti oikeaa. Havaintomäärän ollessa 10000, residuaalimatriisi voi näyttää tällaiselta:

```
 26  1 SURVO 84C EDITOR Mon May 16 08:23:43 1994          D:\M\MEN\ 300 100 0
180 *.....................................................................
181 */MNSIMUL R,*,FRANCIS5,10000 / RND=rand(199401)
182 *CORR FRANCIS5
183 *FACTA CORR.M,3
184 */TRAN-SYMMETR FACT.M,F
185 *MAT LOAD L.M,##.###,END+2_ / Transformation matrix
186 *MAT LOAD E.M,##.###,END+2  / Residual matrix
187 *
188 *MATRIX E.M
189 *Residual_matrix
190 *///          F1     F2     F3
191 *X1         0.004 -0.005 -0.008
192 *X2        -0.004  0.010 -0.004
193 *X3         0.003  0.017 -0.028
194 *X4        -0.017  0.005 -0.013
195 *X5         0.006  0.002 -0.003
196 *X6         0.010 -0.023  0.064
197 *X7         0.006 -0.005  0.020
198 *X8        -0.034  0.052 -0.038
199 *X9         0.016 -0.008 -0.012
200 *X10        0.013 -0.046  0.026
201 *
```

Ei siis ole epäilystäkään siitä, etteikö faktorianalyysi tässä patologisessa tilanteessa antaisi oikeita tuloksia. Havaintomäärän vain tulee olla poikkeuksellisen suuri, jotta kahden jälkimmäisen faktorin heikot signaalit erottuisivat ympärillä olevasta kohinasta.

Samanlaiset jatkotarkastelut tehoavat Francisin muihinkin esimerkkeihin, jopa sellaisiin, joissa faktorimatriisit ovat vajaa-asteisia. Niissä tapauksissa "oikeiden" rakenteiden löytyminen automaattisesti on yhtä mahdotonta kuin "oikeiden" regressiokertoimien saaminen regressiomallista, jossa selittävien muuttujien välillä on täsmällisiä lineaarisia riippuvuuksia.

On selvää, että monissa sovelluksissa oikean faktorien lukumäärän löytäminen voi olla hankalaa. Kun näin tapahtuu, se osoittaa, ettei aineisto kunnolla täytä faktorianalyysin vaatimuksia tai otoskoko on liian pieni. Vastaavat ongelmat koskevat monia muitakin tilastollisia menetelmiä. Faktorianalyysi ei ole mikään poikkeus.

6. Kanoniset korrelaatiot

6.1 Määritelmä

Tarkastellaan samaan aineistoon kuuluvan kahden eri muuttujaryhmän välisiä riippuvuuksia. Tätä aihetta sivuttiin jo aikaisemmin multinormaalijakauman yhteiskorrelaatiokertoimia käsiteltäessä. Tarkoituksena on löytää kummastakin muuttujaryhmästä sellaiset painotetut summat, joiden väliset korrelaatiokertoimet ovat mahdollisimman suuria. Kun toisessa ryhmässä on vain yksi muuttuja, palaudutaan yhteiskorrelaatiokertoimeen. Nyt tutkitaan yleistä tilannetta.

Kuten aikaisemminkin, p muuttujan satunnaisvektori $\mathbf{X}$ ositetaan kahtia niin, että ensimmäisessä osassa $\mathbf{X}^{(1)}$ on q ensimmäistä muuttujaa ja jälkimmäisessä $\mathbf{X}^{(2)}$ loput p-q muuttujaa. Toistaiseksi multinormaalisuus ei ole välttämätöntä, mutta oletamme kuitenkin, että muuttujien yhteisjakaumalla on odotusarvovektori μ ja kovarianssimatriisi Σ, jotka ositetaan vastaavasti merkinnöin

$$\mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11}\ \Sigma_{12} \\ \Sigma_{21}\ \Sigma_{22} \end{bmatrix}.$$

Tehtävänä on löytää muuttujaryhmistä yhdistetyt muuttujat

$$\alpha'\mathbf{X}^{(1)} = \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_q X_q ,$$

$$\beta'\mathbf{X}^{(2)} = \beta_1 X_{q+1} + \beta_2 X_{q+2} + ... + \beta_p X_p ,$$

joiden välinen korrelaatiokerroin on mahdollisimman suuri.

Oletamme, että α- ja β-kertoimet on normeerattu siten, että yhdistettyjen muuttujien varianssit ovat ykkösiä. Koska odotusarvoilla ei ole vaikutusta tarkasteluun, voimme myös olettaa, että $\mu=\mathbf{0}$.

Tällöin korrelaatiokertoimen maksimointi tarkoittaa lausekkeen

$$\mathrm{cov}(\alpha'\mathbf{X}^{(1)},\beta'\mathbf{X}^{(2)}) = \mathrm{E}(\alpha'\mathbf{X}^{(1)}\mathbf{X}^{(2)}{}'\beta) = \alpha'\Sigma_{12}\beta$$

maksimointia kerroinvektoreiden α ja β suhteen ehdoilla

$$\mathrm{var}(\alpha'\mathbf{X}^{(1)}) = \mathrm{E}(\alpha'\mathbf{X}^{(1)}\mathbf{X}^{(1)}{}'\alpha) = \alpha'\Sigma_{11}\alpha = 1 ,$$

$$\mathrm{var}(\beta'\mathbf{X}^{(2)}) = \mathrm{E}(\beta'\mathbf{X}^{(2)}\mathbf{X}^{(2)}{}'\beta) = \beta'\Sigma_{22}\beta = 1 .$$

Otamme käyttöön Cholesky-hajotelmat

$$\Sigma_{11}=\mathbf{S}_1'\mathbf{S}_1 \text{ ja } \Sigma_{22}=\mathbf{S}_2'\mathbf{S}_2$$

sekä uudet vektorit $\mathbf{u}=\mathbf{S}_1\alpha$ ja $\mathbf{v}=\mathbf{S}_2\beta$. Tällöin tehtävämme muuntuu lausekkeen

$$\mathbf{u}'(\mathbf{S}_1^{-1})'\Sigma_{12}\mathbf{S}_2^{-1}\mathbf{v} = \mathbf{u}'\mathbf{A}\mathbf{v}$$

maksimoinniksi ehdoilla $\mathbf{u}'\mathbf{u}=\mathbf{v}'\mathbf{v}=1$. Tässä on merkitty lyhyyden vuoksi

$$\mathbf{A} = (\mathbf{S}_1^{-1})'\Sigma_{12}\mathbf{S}_2^{-1}$$

ja $\mathbf{A}$ on muodoltaan $q\times(p\text{-}q)$-matriisi.

Yleisesti $\mathbf{u}'\mathbf{A}\mathbf{v}$ maksimoituu ehdoilla $\mathbf{u}'\mathbf{u}=\mathbf{v}'\mathbf{v}=1$, kun matriisin $\mathbf{A}$ singulaariarvohajotelmasta $\mathbf{A}=\mathbf{U}\mathbf{D}\mathbf{V}'$ valitaan suurin singulaariarvo d_1 sekä tätä vastaavat (ensimmäiset) pystyvektorit $\mathbf{u}^{(1)}$ ja $\mathbf{v}^{(1)}$ ortogonaalisista matriiseista $\mathbf{U}$ ja $\mathbf{V}$. Singulaariarvo d_1 on samalla lausekkeen $\mathbf{u}'\mathbf{A}\mathbf{v}$ maksimiarvo eli suurin mahdollinen korrelaatiokerroin. Sitä sanotaan ensimmäiseksi kanoniseksi korrelaatiokertoimeksi. Koska $\mathbf{u}=\mathbf{S}_1\alpha$ ja $\mathbf{v}=\mathbf{S}_2\beta$, tämän korrelaatiokertoimen antavat yhdistetyt muuttujat

$$Y_1^{(1)} = \alpha'\mathbf{X}^{(1)} = \mathbf{u}^{(1)'}(\mathbf{S}_1^{-1})'\mathbf{X}^{(1)} ,$$

$$Y_1^{(2)} = \beta'\mathbf{X}^{(2)} = \mathbf{v}^{(1)'}(\mathbf{S}_2^{-1})'\mathbf{X}^{(2)} .$$

Tämä on ensimmäinen kanoninen muuttujapari.

Yksi kanoninen muuttujapari ei yleensä riitä kuvaamaan muuttujaryhmien riippuvuuksia vaan voidaan muodostaa uusia pareja lisäehdolla, etteivät ne korreloi aikaisempien muuttujaparien kanssa.

Voimakkuudeltaan i. kanoninen muuttujapari on tällöin

$$Y_i^{(1)} = \mathbf{u}^{(i)'}(\mathbf{S}_1^{-1})'\mathbf{X}^{(1)} ,$$

$$Y_i^{(2)} = \mathbf{v}^{(i)'}(\mathbf{S}_2^{-1})'\mathbf{X}^{(2)}$$

ja se antaa suurimman mahdollisen korrelaatiokertoimen, joka on d_i , ehdoilla

$$\rho(Y_i^{(1)},Y_j^{(1)}) = 0 ,$$

$$\rho(Y_i^{(1)},Y_j^{(2)}) = 0 ,$$

$$\rho(Y_i^{(2)},Y_j^{(2)}) = 0 ,$$

$$\rho(Y_i^{(2)},Y_j^{(1)}) = 0 , \; j = 1,2,..., i\text{-}1 .$$

Kanonisten korrelaatiokertoimien ja muuttujaparien suurin mahdollinen määrä on $\min(q, p\text{-}q)$, koska tämä on edellä olevan $q\times(p\text{-}q)$-matriisin $\mathbf{A}$ maksimaalinen aste. Sovelluksissa ensimmäinen pari on usein hyvin ilmeinen, mutta seuraavat saattavat olla mielenkiintoisempia.

Yhdistämällä ylläolevat esitykset, kanoniset muuttujavektorit $\mathbf{Y}^{(1)}$ ja $\mathbf{Y}^{(2)}$ voidaan kirjoittaa muodossa

$$\mathbf{Y}^{(1)} = \mathbf{A}^{(1)'}\mathbf{X}^{(1)} ,$$

$$\mathbf{Y}^{(2)} = \mathbf{A}^{(2)'}\mathbf{X}^{(2)} ,$$

missä kerroinmatriisit $\mathbf{A}^{(1)}$ ja $\mathbf{A}^{(2)}$ ovat

$$\mathbf{A}^{(1)} = \mathbf{S}_1^{-1}\mathbf{U} ,$$

$$\mathbf{A}^{(2)} = \mathbf{S}_2^{-1}\mathbf{V} .$$

Siinä erikoistapauksessa, että q=1, on kuten multinormaalijakauman ominaisuuksia tarkasteltaessa (2.2.5) todettiin

$$d_1 = R_{1.23,\ldots,p}$$

eli ainoa kanoninen korrelaatiokerroin on sama kuin yhteiskorrelaatiokerroin.

Kanoniset korrelaatiokertoimet ovat invariantteja toisaalta muuttujien $\mathbf{X}^{(1)}$ ja toisaalta muuttujien $\mathbf{X}^{(2)}$ sisäisissä, säännöllisissä lineaarimuunnoksissa. Täten tarkastelun lähtökohdaksi voidaan valita korrelaatiomatriisi $\mathbf{P}$ ositettuna samoin kuin $\boldsymbol{\Sigma}$ edellä. Kun siis sovitaan, että $\boldsymbol{\Sigma}=\mathbf{P}$, alkuperäisten muuttujien ja kanonisten muuttujien väliset korrelaatiomatriisit ovat

$$\mathbf{P}^{(1)} = \rho(\mathbf{X}^{(1)},\mathbf{Y}^{(1)}) = E(\mathbf{X}^{(1)}\mathbf{Y}^{(1)'}) = E(\mathbf{X}^{(1)}\mathbf{X}^{(1)'}\mathbf{A}^{(1)}) = \mathbf{P}_{11}\mathbf{A}^{(1)} ,$$

$$\mathbf{P}^{(2)} = \rho(\mathbf{X}^{(2)},\mathbf{Y}^{(2)}) = E(\mathbf{X}^{(2)}\mathbf{Y}^{(2)'}) = E(\mathbf{X}^{(2)}\mathbf{X}^{(2)'}\mathbf{A}^{(2)}) = \mathbf{P}_{22}\mathbf{A}^{(2)} .$$

Kun otetaan huomioon, että

$$\mathbf{P}_{11} = \mathbf{S}_1'\mathbf{S}_1 \quad , \quad \mathbf{A}^{(1)} = \mathbf{S}_1^{-1}\mathbf{U} ,$$

$$\mathbf{P}_{22} = \mathbf{S}_2'\mathbf{S}_2 \quad , \quad \mathbf{A}^{(2)} = \mathbf{S}_2^{-1}\mathbf{V} ,$$

saadaan näille korrelaatiomatriiseille myös esitykset

$$\mathbf{P}^{(1)} = \mathbf{S}_1'\mathbf{U} ,$$

$$\mathbf{P}^{(2)} = \mathbf{S}_2'\mathbf{V} .$$

Kanonisten muuttujavektorien $\mathbf{Y}^{(1)}$ ja $\mathbf{Y}^{(2)}$ välinen korrelaatiomatriisi on

$$\rho(\mathbf{Y}^{(1)},\mathbf{Y}^{(2)}) = E(\mathbf{A}^{(1)'}\mathbf{X}^{(1)}\mathbf{X}^{(2)'}\mathbf{A}^{(2)})$$

$$= \mathbf{A}^{(1)'}\boldsymbol{\Sigma}_{12}\mathbf{A}^{(2)} = \mathbf{U}'(\mathbf{S}_1^{-1})'\boldsymbol{\Sigma}_{12}\mathbf{S}_2^{-1}\mathbf{V} = \mathbf{U}'\mathbf{A}\mathbf{V} = \mathbf{U}'\mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{V} = \mathbf{D}$$

eli kanonisten korrelaatiokertoimien muodostama lävistäjämatriisi, kuten pitääkin.

6.1.1 Esimerkki

Ennen kuin tarkastelemme kanonisten korrelaatioiden ja muuttujien estimointia, käymme läpi teoreettisen esimerkin, joka näyttää, miten edellä esitetyt kaavat toimivat.

```
  1  1 SURVO 84C EDITOR Sun Apr 10 13:28:23 1994        C:\M\MEN2\ 200 100 0
 1 *
 2 *MATRIX R0
 3 *///  X11  X12  X13  X21  X22
 4 *X11  1    0    0    0.9  0
 5 *X12  0    1    0    0    0.5
 6 *X13  0    0    1    0    0
 7 *X21  0.9  0    0    1    0
 8 *X22  0    0.5  0    0    1
 9 *
10 *MATRIX T
11 *///  X11  X12  X13  X21  X22
12 *X11  5    1    5    0    0
13 *X12  6    2   -2    0    0
14 *X13  1    1    3    0    0
15 *X21  0    0    0    4    3
16 *X22  0    0    0   -1    2
17 *_
```

Lähdemme liikkeelle 5×5-korrelaatiomatriisista P_0 (R0) , jonka rakenne on hyvin yksinkertainen. Ainoat riippuvuudet esiintyvät muuttujien X_{11} ja X_{21} välillä (korrelaatiokerroin 0.9) sekä muuttujien X_{12} ja X_{22} välillä (0.5). On siis suoraan nähtävissä, että muuttujaryhmien (X_{11},X_{12},X_{13}) ja (X_{21},X_{22}) kanoniset korrelaatiokertoimet ovat 0.9 ja 0.5 .

Tarkoitus on häiritä tätä selkeää tilannetta niin, että teemme kummassakin muuttujaryhmässä mielivaltaisen, säännöllisen lineaarisen muunnoksen. Tätä varten on valittu muunnosmatriisi **T**, jossa $\mathbf{T}_{12}=\mathbf{0}$ ja $\mathbf{T}_{21}=\mathbf{0}$, jolloin koko 5 muuttujan vektoriin **X** kohdistettu muunnos **TX** ei vaikuta muuttujaryhmien välisiin riippuvuuksiin. Muunnetun satunnaisvektorin **TX** kovarianssimatriisi on tällöin $\mathbf{TP}_0\mathbf{T}'$, joka voidaan normalisoida lopulliseksi korrelaatiomatriisiksi P . Kaikki tämä tapahtuu seuraavilla Survon matriisikäskyillä ja muunnettu korrelaatiomatriisi P talletetaan matriisitiedostoksi R.

```
  1  1 SURVO 84C EDITOR Sun Apr 10 13:56:39 1994        C:\M\MEN2\ 200 100 0
17 *
18 *MAT SAVE R0
19 *MAT SAVE T
20 *MAT R=T*R0          / *R~T*R0 5*5
21 *MAT R=MMT2(R,T)     / *R~T*R0*T' S5*5
22 *MAT D=VD(R)         / *D~VD(T*R0*T') 5*1
23 *MAT TRANSFORM D BY 1/sqrt(X#)
24 *MAT D!=DV(D)         / *D~DV(T(D_by_1/sqrt(X#))) D5*5
25 *MAT R=D*R           / *R~D*T*R0*T' 5*5
26 *MAT R!=R*D           / *R~D*T*R0*T'*D 5*5
27 *MAT LOAD R,CUR+2
28 *_
29 *MATRIX R
30 *///           X11      X12      X13      X21      X22
31 *X11       1.00000  0.46442  0.88662  0.54611 -0.21918
32 *X12       0.46442  1.00000  0.09091  0.74172 -0.22923
33 *X13       0.88662  0.09091  1.00000  0.30754  0.01348
34 *X21       0.54611  0.74172  0.30754  1.00000  0.17889
35 *X22      -0.21918 -0.22923  0.01348  0.17889  1.00000
36 *
```

Nuo muuttujaryhmien sisäiset muunnokset ovat sotkeneet alkuperäisen kauniin rakenteen. Suurin suoraan havaittava korrelaatiokerroin muuttujaryhmien välillä on vain 0.74 .

Laskemme nyt matriisitulkin avulla matriisista R lähtien kanoniset korrelaatiot seuraavasti

```
  1  1 SURVO 84C EDITOR Sun Apr 10 14:05:30 1994        C:\M\MEN2\ 200 100 0
 36 *
 37 *MAT R11!=R(1:3,1:3)
 38 *MAT R22!=R(4:5,4:5)
 39 *MAT R12!=R(1:3,4:5)
 40 *MAT S1=CHOL(R11)   / *S1~CHOL(R11) 3*3
 41 *MAT S1!=S1'        / *S1~CHOL(R11)' 3*3
 42 *MAT S2=CHOL(R22)   / *S2~CHOL(R22) 2*2
 43 *MAT S2!=S2'        / *S2~CHOL(R22)' 2*2
 44 *MAT SI1=INV(S1)    / *SI1~INV(S1) 3*3
 45 *MAT SI2=INV(S2)    / *SI2~INV(S2) 2*2
 46 *MAT A=MTM2(SI1,R12)          / *A~INV(S1)'*R12 3*2
 47 *MAT A=A*SI2        / *A~INV(S1)'*R12*INV(S2) 3*2
 48 *MAT SINGULAR_VALUE DECOMPOSITION OF A TO U,D,V
 49 *MAT LOAD D,CUR+2
 50 *
 51 *MATRIX D
 52 *Dsvd(INV(S1)'*R12*INV(S2))
 53 *///      sing.val
 54 *svd1     0.900000
 55 *svd2     0.500000
 56 *
```

ja näemme, että tulokset täsmäävät odotettuihin.

6.2 Kanonisten korrelaatioiden estimointi

Kanoniset korrelaatiot lasketaan analogisesti edellä esitettyjen kaavojen mukaan lähtien aineistosta lasketusta korrelaatiomatriisista **R** . Jos kysymyksessä on otos multinormaalijakaumasta, tulokset ovat suurimman uskottavuuden estimaatteja.

Survossa tehtävä kannattaa suorittaa sukrolla /CANCORR . Jos esim. haluamme tutkia kymmenotteluaineistossa ensimmäisen ja toisen päivän lajien riippuvuuksia, aktivoimme viisi ensimmäistä lajimuuttujaa kirjaimella X ja loput viisi lajimuuttujaa kirjaimella Y . Tässä muuttujien valinta osoitetaan lyhyesti MASK-täsmennyksellä ja /CANCORR-komento käynnistetään muodossa:

```
 22  1 SURVO 84C EDITOR Sun Apr 10 14:29:03 1994        C:\M\MEN2\ 100 100 0
  1 *
  2 *MASK=--XXXXXYYYYY
  3 */CANCORR KYMMEN,CUR+1
  4 *
```

Tulokset ilmestyvät riveille 4-29:

```
  1  1 SURVO 84C EDITOR Sun Apr 10 14:29:17 1994        C:\M\MEN2\ 100 100 0
  1 *
  2 *MASK=--XXXXXYYYYY
  3 */CANCORR KYMMEN,CUR+1
  4 *Canonical analysis on KYMMEN:
  5 *Correlation    CHI^2     P       df      (LCAN.M)
  6 * 1   0.7894   62.737  0.99996    25
  7 * 2   0.5284   22.231  0.86413    16
  8 * 3   0.3306    8.6421 0.52905     9
  9 * 4   0.2823    3.8400 0.57191     4
 10 * 5   0.0972    0.3941 0.46987     1
 11 *Coefficients for canonical variables saved in XCOEFF.M,YCOEFF.M
 12 *
 13 *MATRIX XCAN.M
 14 *Correlations_of_can.variables_with_X
 15 *///        CAN1   CAN2   CAN3   CAN4   CAN5
 16 *M100     -0.218  0.729  0.124  0.633 -0.069
 17 *Pituush  -0.124  0.566 -0.731 -0.301 -0.200
 18 *Kuula    -0.892 -0.400 -0.130  0.121 -0.112
 19 *Korkeus  -0.250 -0.168 -0.140 -0.424  0.843
 20 *M400      0.424  0.174 -0.453  0.762  0.067
 21 *
 22 *MATRIX YCAN.M
 23 *Correlations_of_can.variables_with_Y
 24 *///        CAN1   CAN2   CAN3   CAN4   CAN5
 25 *Aidat    -0.226  0.555 -0.627  0.496 -0.045
 26 *Kiekko   -0.928 -0.336 -0.090 -0.075 -0.108
 27 *Seiväs    0.191  0.299  0.193 -0.232 -0.885
 28 *Keihäs   -0.000 -0.139 -0.540 -0.808  0.192
 29 *M1500     0.806 -0.497 -0.166  0.251 -0.111
 30 *
```

Kanonisten korrelaatioiden ohella /CANCORR poimii esiin alkuperäisten ja kanonisten muuttujien väliset korrelaatiomatriisit, joista on ehkä helpoin päätellä, millä tavalla muuttujaryhmät ovat tekemisissä keskenään.

Tässä tapauksessa ensimmäinen kanoninen muuttujapari liittyy "voimaan" ja toinen "nopeuteen". On kuitenkin syytä kiinnittää huomiota kanonisten korrelaatiokertoimien tilastolliseen merkitsevyyteen, josta tulostuksessa annetaan asymptoottiset χ^2-testisuureet vapausasteineen ja *P*-arvoineen kunkin kanonisen korrelaatiokertoimen osalta (rivit 6-10).

Testisuure, jota tässä käytetään sen testaamiseen, että kanoniset korrelaatiot kertaluvun *k* jälkeen eivät ole nollasta poikkeavia on

$$-[N - 1 - (p + 1)/2] \sum_{i=k+1}^{s} \log(1 - d_i^2) ,$$

missä *N* on otoskoko ja $s=\min(q, p-q)$ kanonisten korrelaatiokertoimien lukumäärä. Arvolla $k=0$ kyseessä on karkea testi muuttujaryhmien riippumattomuudelle.

Koska testin *P*-arvo esimerkissämme arvolla $k=1$ on 0.86, kanoniset korrelaatiokertoimet eivät siis ensimmäistä lukuunottamatta näytä merkitseviltä tässä tapauksessa.

6.3 Informaatioteoreettinen tulkinta

Tässä lyhyessä katsauksessa esittelemme multinormaalijakauman informaatioteoreettisia ominaisuuksia ja näytämme niiden yhteydet kanonisiin korrelaatiokertoimiin.

Entropiamitta
Jatkuvan p-ulotteisen satunnaismuuttujan $\mathbf{X}$ entropiaksi määritellään

$$H(\mathbf{X}) = -\int_{R^p} f(\mathbf{x}) \log f(\mathbf{x})\, \mathbf{dx} .$$

Se kuvaa jakaumaan liittyvää epävarmuutta ja vastaa diskreetin jakauman entropiaa

$$-\sum_{i=1}^{n} p_i \log_2 p_i ,$$

joka kertoo, kuinka monta "on-ei"-tyyppistä kysymystä vastaavan satunnaiskokeen tuloksesta on keskimäärin esitettävä, jotta saadaan selville toteutunut vaihtoehto. Diskreetissä tapauksessa logaritmin kantaluku 2 valitaan tuon "bittitulkinnan" vuoksi. Jatkuvassa tilanteessa vastaava tulkinta on mahdoton, jolloin on yksinkertaisinta käyttää luonnollisia logaritmeja.

Multinormaalijakauman $N(\mu,\Sigma)$ entropia johdetaan seuraavasti. Tässä tapauksessa on

$$\log f(\mathbf{x}) = -\tfrac{1}{2}\,[p \log 2\pi + \log|\Sigma| + (\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)] ,$$

jolloin entropiaksi saadaan

$$\begin{aligned} H(\mathbf{X}) &= \tfrac{1}{2}\,\{\, p \log 2\pi + \log|\Sigma| + \mathrm{tr}[\,\Sigma^{-1}\int_{R^p} (\mathbf{x}-\mu)(\mathbf{x}-\mu)' f(\mathbf{x})\, \mathbf{dx}\,]\,\} \\ &= \tfrac{1}{2}\,\{\, p \log 2\pi + \log|\Sigma| + \mathrm{tr}[\Sigma^{-1}\Sigma]\,\} \\ &= \tfrac{1}{2}\,(\, p \log 2\pi + \log|\Sigma| + p\,) . \end{aligned}$$

Voidaan osoittaa (*Rao*, 1965, s.449-450), että niiden jakaumien joukossa, joilla on annettu odotusarvovektori μ ja kovarianssimatriisi Σ, juuri (multi)normaalijakaumalla on suurin entropia.

Informaatiomitta
Muuttujien $\mathbf{X}^{(2)}$ antama informaatio muuttujista $\mathbf{X}^{(1)}$ määritellään muodossa

$$I(\mathbf{X}^{(1)},\mathbf{X}^{(2)}) = H(\mathbf{X}^{(1)}) - H(\mathbf{X}^{(1)} \mid \mathbf{X}^{(2)})$$

eli se kertoo, kuinka paljon muuttujiin $\mathbf{X}^{(1)}$ liittyvä entropia vähenee, jos saadaan tietää, mitä $\mathbf{X}^{(2)}$ on.

Multinormaalijakaumassa $N(\mu,\Sigma)$ pätee tällöin yksinkertaisesti

$$I(\mathbf{X}^{(1)},\mathbf{X}^{(2)}) = \tfrac{1}{2}\,(\log|\Sigma_{11}| - \log|\Sigma_{11.2}|)\ ,$$

mikä yhtälön

$$|\Sigma| = |\Sigma_{22}||\Sigma_{11.2}|$$

perusteella voidaan kirjoittaa myös symmetrisessä muodossa

$$I(\mathbf{X}^{(1)},\mathbf{X}^{(2)}) = \tfrac{1}{2}\log\frac{|\Sigma_{11}||\Sigma_{22}|}{|\Sigma|}$$

eli $I(\mathbf{X}^{(1)},\mathbf{X}^{(2)}) = I(\mathbf{X}^{(2)},\mathbf{X}^{(1)})$.

Osoitamme nyt, että $I(\mathbf{X}^{(1)},\mathbf{X}^{(2)})$ voidaan esittää kanonisten korrelaatiokertoimien avulla. Teemme täsmentävän oletuksen, että $q \le p\text{-}q$, jolloin kanonisten korrelaatiokertoimien lukumäärä s on q ja $q\times(p\text{-}q)$-matriisin

$$\mathbf{A} = (\mathbf{S}_1^{-1})'\Sigma_{12}\mathbf{S}_2^{-1}$$

singulaariarvohajotelmassa $\mathbf{A=UDV}'$ matriisi $\mathbf{U}$ on ortogonaalinen. Käytämme lisäksi hyväksemme aikaisempia hajotelmia

$$\Sigma_{11} = \mathbf{S}_1'\mathbf{S}_1\ ,$$
$$\Sigma_{22} = \mathbf{S}_2'\mathbf{S}_2\ ,$$

jolloin voimme kehitellä informaatiomittaa $I(\mathbf{X}^{(1)},\mathbf{X}^{(2)})$ seuraavasti:

$$I(\mathbf{X}^{(1)},\mathbf{X}^{(2)}) = \tfrac{1}{2}\,(\log|\Sigma_{11}| - \log|\Sigma_{11.2}|)$$
$$= \tfrac{1}{2}\,(\log|\Sigma_{11}| - \log|\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|)$$
$$= -\tfrac{1}{2}\log[|\Sigma_{11}^{-1}|\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|]$$
$$= -\tfrac{1}{2}\log[|(\mathbf{S}_1^{-1})'||\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}||\mathbf{S}_1^{-1}|]$$
$$= -\tfrac{1}{2}\log|\mathbf{I} - (\mathbf{S}_1^{-1})'\Sigma_{12}\mathbf{S}_2^{-1}(\mathbf{S}_2^{-1})'\Sigma_{21}\mathbf{S}_1^{-1}|$$
$$= -\tfrac{1}{2}\log|\mathbf{I} - \mathbf{AA}'|$$
$$= -\tfrac{1}{2}\log|\mathbf{I} - \mathbf{UDV}'\mathbf{VDU}'|$$
$$= -\tfrac{1}{2}\log|\mathbf{I} - \mathbf{UD}^2\mathbf{U}'|$$
$$= -\tfrac{1}{2}\log(|\mathbf{U}'||\mathbf{I} - \mathbf{UD}^2\mathbf{U}'||\mathbf{U}|)$$
$$= -\tfrac{1}{2}\log|\mathbf{I} - \mathbf{D}^2|$$
$$= -\tfrac{1}{2}\sum_{i=1}^{s}\log(1 - d_i^2)\ .$$

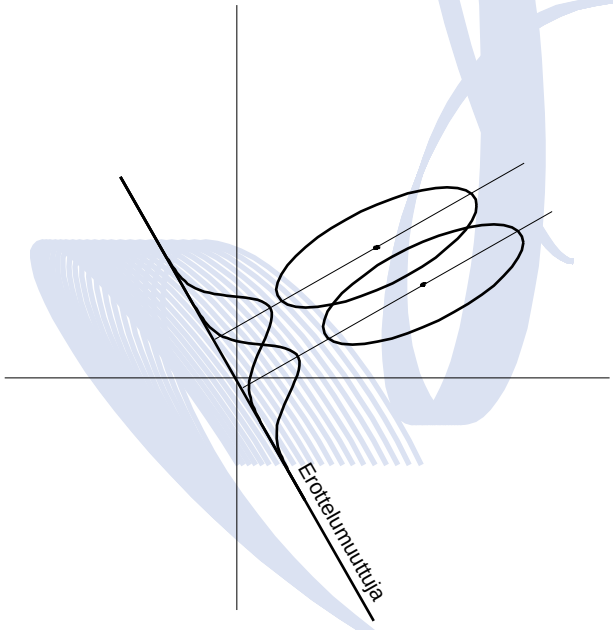
$I(\mathbf{X}^{(1)},\mathbf{X}^{(2)})$ on siis vakiotekijää vaille edellä mainittu muuttujaryhmien riippumattomutta koskeva testisuure.

7. Erotteluanalyysi

7.1 Määritelmä

Tarkastellaan satunnaisvektoria **X** *g* eri perusjoukossa, joista kustakin on käytettävissä otos. Tehtävänä on määrätä sellaiset yhdistetyt muuttujat, jotka parhaiten kuvaavat perusjoukkojen (ryhmien) välisiä eroja.

Alustavana esimerkkinä vertailemme jakaumia $N(\mu^{(1)},\Sigma)$ ja $N(\mu^{(2)},\Sigma)$, joilla on siis sama kovarianssimatriisi mutta eri odotusarvovektorit. Määräämme sen yhdistetyn muuttujan **a**'**X** , joka erottaa jakaumat selvimmin toisistaan. Tällaista muuttujaa sanotaan *erottelumuuttujaksi* (diskriminaattoriksi). Vektori **a** määrätään siis suuntana, johon projisoidut jakaumat erottuvat parhaiten. Esim. kahden muuttujan tapauksessa erottelumuuttujan suunta voisi näyttää seuraavalta.



Ns. Fisherin erottelumuuttuja määräytyy siten, että maksimoidaan lauseke

$$[\mathbf{a}'(\mu^{(1)} - \mu^{(2)})]^2$$

vektorin **a** suhteen ehdolla

$$\mathrm{var}(\mathbf{a}'\mathbf{X}) = \mathbf{a}'\Sigma\mathbf{a} = 1 .$$

Soveltamalla Cholesky-hajotelmaa $\Sigma=\mathbf{C}\mathbf{C}'$ ja muunnosta $\mathbf{b}=\mathbf{C}'\mathbf{a}$ maksimointitehtävä tulee muotoon: On maksimoitava **b**:n suhteen

$$\mathbf{b}'\mathbf{C}^{-1}(\mu^{(1)} - \mu^{(2)})(\mu^{(1)} - \mu^{(2)})'(\mathbf{C}^{-1})'\mathbf{b} = \mathbf{b}'\mathbf{u}\mathbf{u}'\mathbf{b} ,$$

missä

$$\mathbf{u} = \mathbf{C}^{-1}(\mu^{(1)} - \mu^{(2)})$$

ehdolla

$$\mathbf{b}'\mathbf{b}=1 .$$

Tehtävä on täysin samanlainen kuin Hotellingin T^2-testin yhteydessä tapaamamme. Optimaalinen **b** on verrannollinen vektoriin **u** eli

$$\mathbf{a} \propto (\mathbf{C}^{-1})'\mathbf{b} = (\mathbf{C}^{-1})'\mathbf{C}^{-1}(\mu^{(1)} - \mu^{(2)}) = \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) .$$

Kun toimitaan teoreettisten parametrien asemasta otosten pohjalta, odotusarvot korvataan otoskeskiarvoilla ja Σ estimoidaan samalla tavalla kuin T^2-testissä kahden otoksen tapauksessa.

Itse asiassa Survon kahden multinormaalisen otoksen vertailuun tarkoitettu sukro `/MTEST-T2/2` laskee kahden otoksen T^2-testin ohella ko. erottelumuuttujan kertoimet ja tallettaa ne matriisitiedostoon `T2COEFF.M` toiseksi sarakkeeksi otsikolla "`discr`". (Ensimmäinen sarake on otoskeskiarvojen erotus.)

Siirrymme nyt yleiseen tilanteeseen, jossa vertailtavien ryhmien määrä g on enemmän kuin 2. Aivan samalla tavalla kuin t-testi yleistettiin T^2-testiksi, erotteluanalyysia voidaan tarkastella 1-suuntaisen varianssianalyysin pohjalta. Kertaamme tämän vuoksi lyhyesti ko. varianssianalyysin päätulokset.

Tarkasteltakoon reaaliarvoista satunnaismuuttujaa Y g eri ryhmässä. Olkoon k. ryhmässä (k=1,2,...,g)

$$Y \sim N(\mu_k,\sigma^2) , \quad k = 1,2,...,g$$

ja tästä ryhmästä saatavilla N_k riippumatonta havaintoa

$$Y_{k\alpha} , \alpha = 1,2,...,N_k .$$

Otamme käyttöön vielä seuraavat merkinnät:

$$\overline{Y}_k = \frac{1}{N_k}\sum_{\alpha=1}^{N_k} Y_{k\alpha} \qquad \text{(keskiarvo } k\text{. ryhmässä)}$$

$$N = N_1 + N_2 + ... + N_g \qquad \text{(havaintojen kokonaismäärä)}$$

$$\overline{Y} = \frac{1}{N}\sum_{k=1}^{g} N_k \overline{Y}_k \qquad \text{(yleiskeskiarvo)}$$

Hypoteesia H_0: $\mu_1=\mu_2=...=\mu_g$ testattaessa päädytään varianssianalyysikaavioon:

Vaihtelu	Neliösumma	Vap.ast.	s^2	F
Ryhmien välinen	$Q_2(Y) = \sum_{k=1}^{g} N_k(\overline{Y}_k - \overline{Y})^2$	$g-1$	s_2^2	s_2^2/s_1^2
Ryhmien sisäinen	$Q_1(Y) = \sum_{k=1}^{g} \sum_{\alpha=1}^{N_k} (Y_{k\alpha} - \overline{Y}_k)^2$	$N-g$	s_1^2	
Kokonaisvaihtelu	$Q(Y) = \sum_{k=1}^{g} \sum_{\alpha=1}^{N_k} (Y_{k\alpha} - \overline{Y})^2$	$N-1$		

Mikäli H_0 pätee, s_2^2/s_1^2 noudattaa F-jakaumaa vapausastein g-1, N-g .

Tulemme nyt soveltamaan tätä kaaviota yleistetyssä tilanteessa, jossa Y on multinormaalijakaumaa noudattavan satunnaisvektorin $\mathbf{X}$ komponenttien painotettu summa.

Olkoon siis k. ryhmässä

$$\mathbf{X}=(X_1,X_2,...,X_p) \sim N(\mu^{(k)},\Sigma)$$

ja otamme käyttöön merkinnät

$X_{ik\alpha} = X_i$:n arvo k. ryhmän α. havainnossa,

$\mathbf{X}_{k\alpha}$ = k. ryhmän α. havaintovektori,

$$\overline{\mathbf{X}}_k = \frac{1}{N_k} \sum_{\alpha=1}^{N_k} \mathbf{X}_{k\alpha} \ , \quad \overline{\mathbf{X}} = \frac{1}{N} \sum_{k=1}^{g} N_k \overline{\mathbf{X}}_k \ ,$$

$$\mathbf{S}_k = \frac{1}{N_k\text{-}1} \sum_{\alpha=1}^{N_k} (\mathbf{X}_{k\alpha} - \overline{\mathbf{X}}_k)(\mathbf{X}_{k\alpha} - \overline{\mathbf{X}}_k)' = k. \text{ ryhmän otoskovarianssimatriisi.}$$

Tarkastellaan mielivaltaista yhdistettyä muuttujaa Y=$\mathbf{a}$'$\mathbf{X}$ ja sovellamme merkintöjä

$$Y_{k\alpha} = \mathbf{a}'\mathbf{X}_{k\alpha} \ , \ \overline{Y}_k = \mathbf{a}'\overline{\mathbf{X}}_k \ , \ \overline{Y} = \mathbf{a}'\overline{\mathbf{X}} \ .$$

Tällöin ryhmien sisäistä vaihtelua kuvaava neliösumma on kirjoitettavissa muodossa

$$Q_1(Y) = \sum_{k=1}^{g} \sum_{\alpha=1}^{N_k} (Y_{k\alpha} - \overline{Y}_k)^2 = \mathbf{a}'\mathbf{W}\mathbf{a} ,$$

missä

$$\mathbf{W} = \sum_{k=1}^{g} \sum_{\alpha=1}^{N_k} (\mathbf{X}_{k\alpha} - \overline{\mathbf{X}}_k)(\mathbf{X}_{k\alpha} - \overline{\mathbf{X}}_k)' = \sum_{k=1}^{g} (N_k - 1)\mathbf{S}_k .$$

Vastaavasti ryhmien välisen vaihtelun neliösumma on

$$Q_2(Y) = \sum_{k=1}^{g} N_k(\overline{Y}_k - \overline{Y})^2 = \mathbf{a}'\mathbf{B}\mathbf{a} ,$$

missä

$$\mathbf{B} = \sum_{k=1}^{g} N_k(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})' ,$$

jolloin yleiskeskiarvosta laskettujen poikkeamien kokonaisneliösummaksi tulee

$$Q(Y) = Q_1(Y) + Q_2(Y) = \sum_{k=1}^{g} \sum_{\alpha=1}^{N_k} (Y_{k\alpha} - \overline{Y})^2 = \mathbf{a}'\mathbf{T}\mathbf{a} ,$$

missä

$$\mathbf{T} = \sum_{k=1}^{g} \sum_{\alpha=1}^{N_k} (\mathbf{X}_{k\alpha} - \overline{\mathbf{X}})(\mathbf{X}_{k\alpha} - \overline{\mathbf{X}})' = \mathbf{W} + \mathbf{B} .$$

Matriisin **B** aste olkoon $m \le \min(g\text{-}1, p)$ ja matriisin **W** maksimaalinen eli p . Tällöin, mikäli H_0: $\mu^{(1)}=\mu^{(2)}=...=\mu^{(g)}$ pätee,

$$s_2^2/s_1^2 = \frac{N-g}{g-1} \cdot \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}}$$

noudattaa F-jakaumaa vapausastein g-1 ja N-g kaikilla (nollavektorista eroavilla) **a**-vektoreilla.

Erottelumuuttujat Y=**a**'**X** muodostetaan maksimoimalla edellä johdettu F-testisuure **a**:n suhteen eli saatetaan testikriteerin mielessä ryhmät mahdollisimman erilleen toisistaan.

Tehtävä voidaan esittää myös muodossa: On maksimoitava

a'**Ba** ehdolla **a**'**Wa**=1

a:n suhteen. Käyttämällä matriisin **W** Cholesky-hajotelmaa

W = **C**$\mathbf{C}'$

ja merkitsemällä

b=**C**'**a**

joudutaan maksimoimaan

b'$\mathbf{C}^{-1}$**B**($\mathbf{C}^{-1}$)'**b** ehdolla **b**'**b**=1

b:n suhteen eli ratkaisu saadaan spektraalihajotelmasta

$$\mathbf{C}^{-1}\mathbf{B}(\mathbf{C}^{-1})' = \mathbf{U}\Lambda\mathbf{U}'$$

muodossa

$$\mathbf{a} = (\mathbf{C}^{-1})'\mathbf{b} = (\mathbf{C}^{-1})'\mathbf{u}^{(1)} ,$$

missä $\mathbf{u}^{(1)}$ on suurinta ominaisarvoa λ_1 vastaava ominaisvektori eli matriisin **U** ensimmäinen sarake.

Matriisin **B** aste *m*, joka on korkeintaan min(*g*-1, *p*), määrää mahdollisten erottelumuuttujien lukumäärän ja erottelumuuttujat ovat

$$Y_i = \mathbf{u}^{(i)\prime}\mathbf{C}^{-1}\mathbf{X} = \mathbf{a}^{(i)\prime}\mathbf{X} , \; i=1,2,...,m .$$

Erottelumuuttujille on ominaista, että Y_i maksimoi em. *F*-testisuureen ehdoilla

$$\mathbf{a}^{(j)\prime}\mathbf{B}\mathbf{a}^{(i)} = \mathbf{a}^{(j)\prime}\mathbf{W}\mathbf{a}^{(i)} = 0 , \; j=1,2,...,i-1 .$$

Nämä ehdot merkitsevät olennaisesti sitä, että

$$r(Y_i,Y_j) = 0 , \text{ kun } i \neq j$$

koko aineistosta laskettuna, sillä

$$(N-1)\text{cov}(Y_i,Y_j) = \mathbf{a}^{(j)\prime}\mathbf{T}\mathbf{a}^{(i)} = \mathbf{a}^{(j)\prime}\mathbf{B}\mathbf{a}^{(i)} + \mathbf{a}^{(j)\prime}\mathbf{W}\mathbf{a}^{(i)} = 0 .$$

Y_i siis maksimoi erottelumuuttujilta $Y_1,Y_2,...,Y_{i-1}$ selittämättä jääneet ryhmien erot.

Kun merkitään

$$\mathbf{A} = [\mathbf{a}^{(1)}\ \mathbf{a}^{(2)}\ ...\ \mathbf{a}^{(m)}] ,$$

on edellisen perusteella

$$\mathbf{A}'\mathbf{B}\mathbf{A} = \Lambda , \; \mathbf{A}'\mathbf{W}\mathbf{A} = \mathbf{I} \text{ ja } \mathbf{A}'\mathbf{T}\mathbf{A} = \Lambda + \mathbf{I} .$$

Pääkomponenttianalyysin tulkinnan tapaan ominaisarvojen $\lambda_1,\lambda_2,...,\lambda_m$ summa kuvaa tällöin ryhmien "kokonaiseroa" ja jokainen ominaisarvo ko. erottelumuuttujan selittämää eroa.

Voidaan osoittaa, että erotteluanalyysi on läheistä sukua kanonisille korrelaatioille. Jos kanoniset korrelaatiot lasketaan aineistosta, jossa erotteluanalyysin muuttujat **X** ovat toisena muuttujaryhmänä ja dikotomiset ryhmien indikaattorimuuttujat toisena muuttujaryhmänä, saadaan kanoniset korrelaatiokertoimet d_i , jotka ovat yhteydessä erotteluanalyysin ominaisarvoihin λ_i kaavan

$$d_i^2 = \lambda_i/(1 + \lambda_i)$$

mukaisesti.

Tällöin analogiseksi testisuureeksi sille, että erottelumuuttujien selittämä ero

k. erottelumuuttujan jälkeen ei ole merkitsevä, voidaan valita

$$[N - 1 - (p + g)/2] \sum_{i=k+1}^{g} \log(1 + \lambda_i) ,$$

joka noudattaa χ^2-jakaumaa $(p\text{-}k)(g\text{-}k\text{-}1)$ vapausasteella, jos eroa ei ole. Arvolla k=0 testataan, onko ryhmien välillä lainkaan keskiarvoeroja eli tämä vastaa 1-suuntaisen varianssianalyysin yleistystä p muuttujalle.

7.2 Luokitteluongelma

Erotteluanalyysiin liittyy läheisesti havaintojen luokittelutehtävä. Oletetaan, että em. erotteluanalyysin tilanteessa saadaan uusia havaintoja **X**, joiden alkuperää, so. todellista ryhmää k=1,2,...,g , ei tunneta. Syntyy ongelma, miten uudet havainnot tulisi tällöin luokitella.

Asiaa voi tutkia joko alkuperäisten muuttujien **X** tai erottelumuuttujien **Y**=**A**'**X** avulla. Jälkimmäinen tapa on tavallisesti järkevämpi, sillä

1) muuttujia on paljon vähemmän,
2) erottelumuuttujat ovat erottelun kannalta karakteristisempia,
3) erottelu- ja luokitteluinformaatio säilyy,
4) multinormaalisuusolettamus on paremmin voimassa.

Luokittelu perustuu Mahalanobis-etäisyyksiin

$$D_k^2 = (\mathbf{X} - \overline{\mathbf{X}}_k)'\mathbf{S}_k^{-1}(\mathbf{X} - \overline{\mathbf{X}}_k) , \quad k = 1,2,...,g ,$$

missä $\mathbf{S}_k$ on k. ryhmästä laskettu otoskovarianssimatriisi. Jos kovarianssit voidaan olettaa ryhmästä riippumatta samoiksi, kuten edellä tehtiin, on perusteltua korvata jokainen $\mathbf{S}_k$ matriisilla

$$\mathbf{S} = \frac{1}{N\text{-}g}\mathbf{W} .$$

Käytettäessä erottelumuuttujia **Y**=**A**'**X** alkuperäisten muuttujien asemasta, ko. etäisyydet ovat muotoa

$$D_k^2 = (\mathbf{Y} - \overline{\mathbf{Y}}_k)'(\mathbf{A}'\mathbf{S}_k\mathbf{A})^{-1}(\mathbf{Y} - \overline{\mathbf{Y}}_k) , \quad k = 1,2,...,g .$$

On olemassa useita erilaisia luokitteluperiaatteita. Näistä tärkeimmät ovat seuraavat:

1) Sijoitetaan havainto **X** siihen ryhmään, jossa Mahalanobis-etäisyys on pienin.

2) Käytetään Bayes-periaatetta eli oletetaan, että kuhunkin ryhmään kuulumiselle on annettu a priori -todennäköisyys p_k , k=1,2,...,g . Usein valitaan $p_k = N_k/N$.

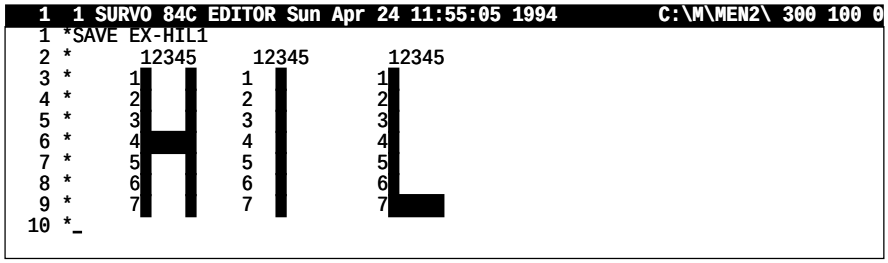
Tällöin havainto **X** sijoitetaan siihen ryhmään k , jolle lauseke

$$p_k|\mathbf{S}_k|^{-1/2}\exp(-\tfrac{1}{2}D_k^2)$$

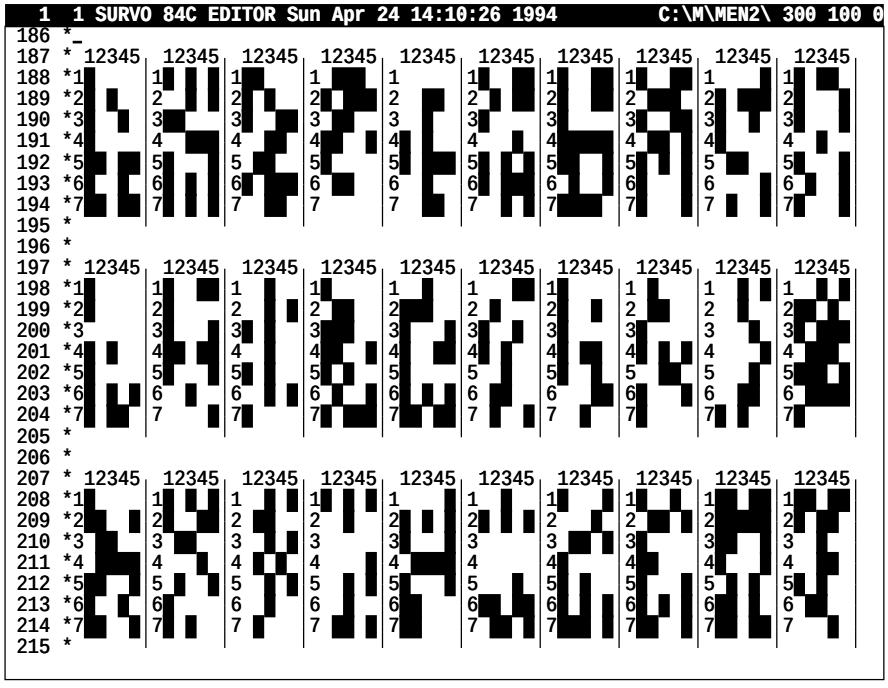
on suurin. Tämä on verrannollinen a posteriori -todennäköisyyteen multinormaalisuuden vallitessa.

7.2.1 Esimerkki hahmontunnistuksesta

Erotteluanalyysin ja siihen liittyvän havaintojen luokittelutehtävän näytteenä tutkimme seuraavanlaista yksinkertaistettua hahmontunnistustilannetta. Luomme 300 havainnon aineiston kirjaimista H,I,L, jotka alunperin on koodattu 7×5-pistematriiseina alla olevan kuvan mukaisesti. Kutakin havaintoa tullaan häiritsemään "kohinalla", joka muuttaa tietyllä todennäköisyydellä mustia pisteitä valkoisiksi ja päinvastoin.



Seuraavassa kuvassa on 30 havainnon näyte tällä tavalla sotketuista merkeistä. Lukija voi yrittää arvioida, mistä kirjaimesta on kussakin tapauksessa kysymys. Myöhemmin kerrotaan erottelu- ja luokitteluanalyysin antama tulos, johon sopii verrata omia käsityksiään.



Tulemme näkemään, että erotteluanalyysi luokittelee nämä tapaukset niiden

epämääräisyydestä huolimatta "oikein".

Kannattaa myös panna merkille, että alkuperäiset 35 (=7×5) muuttujaa ovat 0-1-arvoisia, joten niiden yhteisjakauma on kaukana multinormaalisesta.

Tarvittava aineisto syntyy seuraavasti:

```
 23  1 SURVO 84C EDITOR Sun Apr 24 15:07:37 1994        C:\M\MEN2\ 300 100 0
10 *
11 *FILE CREATE HIL,128,60_
12 *Kirjainten H,I,L bittikartat
13 *FIELDS:
14 * 1 N 1 K
15 * 2 N 1 X11
16 * 3 N 1 X21
17 * 4 N 1 X31
18 * 5 N 1 X41
19 * 6 N 1 X51
20 * 7 N 1 X61
21 * 8 N 1 X71
22 * 9 N 1 X12
23 *10 N 1 X22
24 *11 N 1 X32
.. ... . . ...
47 *34 N 1 X55
48 *35 N 1 X65
49 *36 N 1 X75
50 *END
51 *
```

Aluksi luodaan Survon havaintotiedosto **HIL** ja siihen ryhmää (kirjainta) osoittava muuttuja **K** sekä kirjaimen bittejä osoittavat muuttujat seuraavan matriisin mukaisesti:

$$
\begin{matrix}
X_{11} & X_{12} & X_{13} & X_{14} & X_{15} \\
X_{21} & X_{22} & X_{23} & X_{24} & X_{25} \\
X_{31} & X_{32} & X_{33} & X_{34} & X_{35} \\
X_{41} & X_{42} & X_{43} & X_{44} & X_{45} \\
X_{51} & X_{52} & X_{53} & X_{54} & X_{55} \\
X_{61} & X_{62} & X_{63} & X_{64} & X_{65} \\
X_{71} & X_{72} & X_{73} & X_{74} & X_{75}
\end{matrix}
$$

Tiedostoon asetetaan 300 tyhjää havaintoa (rivi 52), ne täytetään nollilla (rivi 55) ja muuttujan **K** arvot määrätään siten, että 100 ensimmäistä havaintoa saa arvon 1 (H), 100 seuraavaa arvon 2 (I) ja loput 100 arvon 3 (L):

```
  1  1 SURVO 84C EDITOR Sun Apr 24 15:18:43 1994        C:\M\MEN2\ 300 100 0
51 *
52 *FILE INIT HIL,300
53 *...............................................................................
54 *MASK=-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
55 *TRANSFORM HIL BY 0
56 *...............................................................................
57 *VAR K=1 TO HIL / IND=ORDER,1,100
58 *VAR K=2 TO HIL / IND=ORDER,101,200
59 *VAR K=3 TO HIL / IND=ORDER,201,300
60 *_
```

Kustakin kirjaimesta tehdään 100 havaintoa, jotka vastaavat alussa kuvattua ihannemallia **VAR**-komennoin:

```
  1  1 SURVO 84C EDITOR Sun Apr 24 15:23:33 1994        C:\M\MEN2\ 300 100 0
 61 *.................................................................
 62 *IND=K,1           H-kirjaimet
 63 *X11=1 X21=1 X31=1 X41=1 X51=1 X61=1 X71=1
 64 *X42=1 X43=1 X44=1
 65 *X15=1 X25=1 X35=1 X45=1 X55=1 X65=1 X75=1
 66 *VAR X11,X21,X31,X41,X51,X61,X71 TO HIL
 67 *VAR X42,X43,X44 TO HIL
 68 *VAR X15,X25,X35,X45,X55,X65,X75 TO HIL
 69 *
 70 *.................................................................
 71 *IND=K,2           I-kirjaimet
 72 *X13=1 X23=1 X33=1 X43=1 X53=1 X63=1 X73=1
 73 *VAR X13,X23,X33,X43,X53,X63,X73 TO HIL
 74 *
 75 *.................................................................
 76 *IND=K,3           L-kirjaimet
 77 *X11=1 X21=1 X31=1 X41=1 X51=1 X61=1 X71=1
 78 *X72=1 X73=1 X74=1 X75=1
 79 *VAR X11,X21,X31,X41,X51,X61,X71 TO HIL
 80 *VAR X72,X73,X74,X75 TO HIL
 81 *_
```

Kirjaimiin lisätään voimakas "kohina" muuttamalla jokainen ykkösbitti nollaksi todennäköisyydellä 0.3 ja nollabitti ykköseksi samalla todennäköisyydellä. Tämä tapahtuu suoraan seuraavalla TRANSFORM-komennolla:

```
 21  1 SURVO 84C EDITOR Sun Apr 24 15:27:44 1994        C:\M\MEN2\ 300 100 0
 82 *.................................................................
 83 *Satunnaistaminen:
 84 *Musta piste muuttuu valkoiseksi todennäköisyydellä 0.3 .
 85 *Valkoinen piste muuttuu mustaksi todennäköisyydellä 0.3 .
 86 *MASK=-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
 87 *TRANSFORM HIL BY MIX_
 88 *MIX=if(X=1)then(int(0.7+rand(1001)))else(int(0.3+rand(1001)))
 89 *
```

Tässä on näytteenä kustakin 100 havainnon osaotoksesta ensimmäinen tapaus:

```
  1  1 SURVO 84C EDITOR Sun Apr 24 15:38:43 1994        C:\M\MEN2\ 300 100 0
 89 *.................................................................
 90 *Ensimmäiset havainnot kustakin osaotoksesta:
 91 *
 92 *       H          I          L
 93 *     12345      12345      12345
 94 *     1 ▐▀▌      1▐▌▌       1▐▀▌
 95 *     2▐▀▌▌      2▐▀▌       2▐▀▀▌
 96 *     3▐▀▀▌      3▐▀▌       3▐▀▌
 97 *     4▐▀▐▌      4 ▐▌       4▐▀▌
 98 *     5  ▐▌      5▐▀▀       5▐▀▀
 99 *     6▐▌ ▐▌     6▐▀▌       6▐▌
100 *     7▐▌        7▐▌ ▐▌     7 ▐▀▀
101 *_
```

Varsinainen analyysi alkaa laskemalla osaotoksista keskiarvot, hajonnat ja korrelaatiokertoimet. Ne siirretään matriisitiedostoihin R1,M1,R2,M2,R3,M3:

```
  1  1 SURVO 84C EDITOR Sun Apr 24 16:34:13 1994        C:\M\MEN2\ 300 100 0
102 *.....................................................................
103 *Keskiarvot, hajonnat ja korrelaatiot ryhmittäin:
104 *MASK=-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
105 *CORR HIL / IND=K,1
106 *MAT R1=CORR.M      / *R1~R(HIL) S35*35
107 *MAT M1=MSN.M       / *M1~MSN(HIL) 35*3
108 *CORR HIL / IND=K,2
109 *MAT R2=CORR.M      / *R2~R(HIL) S35*35
110 *MAT M2=MSN.M       / *M2~MSN(HIL) 35*3
111 *CORR HIL / IND=K,3
112 *MAT R3=CORR.M      / *R3~R(HIL) S35*35
113 *MAT M3=MSN.M       / *M3~MSN(HIL) 35*3
114 *_
```

Erotteluanalyysiin käytetään tässä uutta /DISCRI-sukroa. Saman tehtävän toteuttaa Markku Korhosen laatima DISCR-operaatio, joka myös luokittelee havainnot automaattisesti. DISCR-operaatio toimii suoraan havaintoaineiston perusteella ja on aitona C-ohjelmana huomattavasti nopeampi kuin /DISCRI-sukro. Käytämme kuitenkin jälkimmäistä, koska se on kätevämpi opetus- ja kokeilutilanteissa.

/DISCRI-sukro edellyttää, että ryhmittäin on valmiiksi laskettuna korrelaatiomatriisit ja MSN-matriisit. Näihin viitataan CORR- ja MSN-täsmennyksillä ja niistä on helppo rakentaa erotteluanalyysin **W**- ja **B**-matriisit. Itse komento ei tarvitse mitään parametreja. Tuloksena saadaan komennon alapuolelle (seuraavassa rivit 120-130) lyhyt yhteenveto ominaisarvoista, niiden erotteluosuuksista, kanonisista korrelaatioista ja likimääräisestä χ^2-testistä.

```
  1  1 SURVO 84C EDITOR Sun Apr 24 16:38:18 1994        C:\M\MEN2\ 300 100 0
115 *.....................................................................
116 *Erotteluanalyysi:
117 *CORR=R1,R2,R3
118 *MSN=M1,M2,M3
119 */DISCRI
120 *    Eig.val.  %      Can.corr  Chi^2    df  P
121 * 1  3.166677  63.00  0.871780  693.7899  70 0.9999
122 * 2  1.859658  37.00  0.806417  294.1966  34 1
123 *
124 *MAT LOAD DISCRL.M,END+2  / Discriminant coefficients
125 *MAT LOAD DISCRXR.M,END+2 / Correlations variables/discriminators
126 *Correlations, means and standard deviations of discriminators
127 *for each of the 3 groups are saved in matrix files corresponding
128 *to CORR and MSN files with their names preceded by letter 'D'.
129 *Discriminant scores are computed by
130 *LINCO <data_file>,DISCRL.M(D1,D2,...)
131 *_
```

Tässä sovelluksessa molemmat mahdolliset erottelumuuttujat ovat hyvin selvästi merkitseviä.

/DISCRI-sukro tallettaa erottelumuuttujien painokertoimet **A** matriisitiedostoksi DISCRL.M sekä erottelumuuttujien ja alkuperäisten muuttujien väliset korrelaatiokertoimet matriisitiedostoksi DISCRXR.M . Näiden tulostamista varten /DISCRI kirjoittaa valmiit komennot (tässä riveillä 124-125). Painokertoimet **A** talletetaan skaalattuina siten, että **a**'**Sa**=1 , missä **S**=**W**/(*N*-*g*). Tällöin niiden suuruudet eivät riipu havaintojen lukumäärästä.

Lisäksi **/DISCRI** on laskenut ja tallettanut ryhmittäin erottelumuuttujien keskinäiset keskiarvot, hajonnat ja korrelaatiokertoimet matriisitiedostoihin, jotka vastaavat **CORR**- ja **MSN**-täsmennyksissä mainittuja. Nimien eteen on vain lisätty D-kirjain erottamaan ne alkuperäisistä.

Painokerroinmatriisi ja vastaava korrelaatiomatriisi näyttävät rinnakkain aseteltuina seuraavilta:

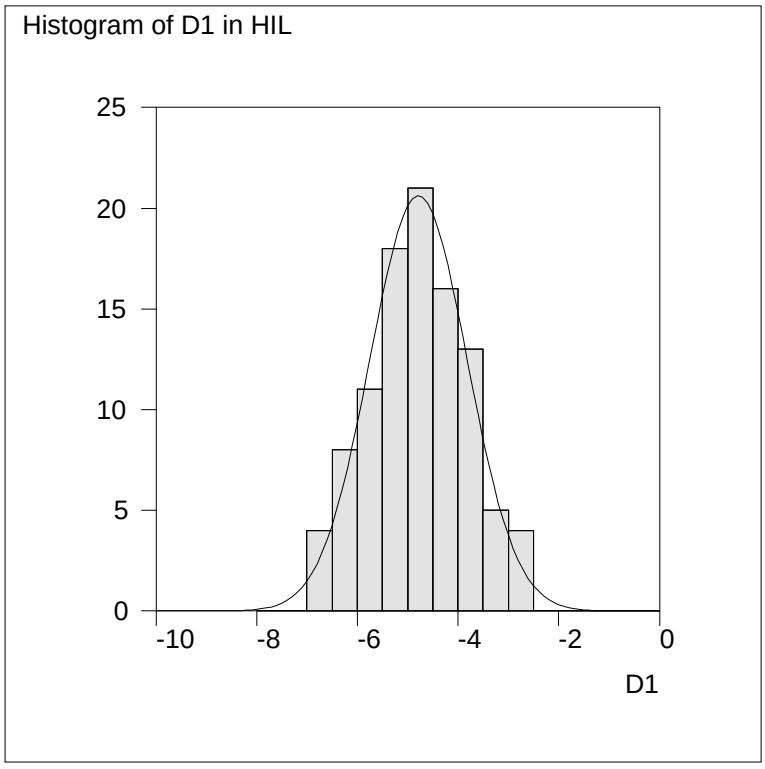
```
  1  1 SURVO 84C EDITOR Sun Apr 24 17:04:21 1994        C:\M\MEN2\ 300 100 0
131 *
132 *Matriisit siirrettynä vertailua varten vierekkäin:
133 *MATRIX DISCRL.M             MATRIX DISCRXR.M
134 *Discriminator_loadings      Correlations_between_variables_and_discrimin
135 *///            %1       %2  ///        Discr1   Discr2
136 *X11      -0.35555  0.24780  X11      -0.38354  0.16640
137 *X21      -0.54086  0.24606  X21      -0.34263  0.13033
138 *X31      -0.45305  0.33676  X31      -0.41697  0.10264
139 *X41      -0.69289  0.06400  X41      -0.56177  0.03394
140 *X51      -0.55529  0.16237  X51      -0.44303  0.06797
141 *X61      -0.81668  0.60068  X61      -0.46883  0.18555
142 *X71      -0.48370  0.05659  X71      -0.36816  0.06457
143 *X12      -0.09118 -0.04808  X12      -0.08029 -0.06371
144 *X22      -0.16507  0.12567  X22      -0.04667  0.07512
145 *X32       0.23745 -0.04745  X32       0.11380 -0.02771
146 *X42      -0.31848 -0.89135  X42      -0.35916 -0.41927
147 *X52      -0.01521  0.29840  X52      -0.04185  0.10664
148 *X62      -0.31208  0.17670  X62      -0.13463  0.08283
149 *X72      -0.12488  0.64358  X72      -0.11760  0.39718
150 *X13       0.51404 -0.27779  X13       0.37493 -0.17416
151 *X23       0.41393  0.03554  X23       0.38348 -0.18736
152 *X33       0.21736 -0.14527  X33       0.29500 -0.09205
153 *X43       0.28752 -0.68820  X43       0.09318 -0.42103
154 *X53       0.46922 -0.37748  X53       0.38948 -0.21111
155 *X63       0.49321 -0.39669  X63       0.43650 -0.21643
156 *X73       0.36013  0.63809  X73       0.24761  0.36625
157 *X14      -0.11975  0.43458  X14      -0.10870  0.08346
158 *X24      -0.12968  0.13070  X24      -0.04206  0.05951
159 *X34       0.25972  0.20172  X34       0.05730  0.02628
160 *X44      -0.38441 -0.22126  X44      -0.34636 -0.25552
161 *X54       0.05037  0.12873  X54      -0.06648  0.01133
162 *X64       0.14475  0.08950  X64      -0.02740  0.03279
163 *X74      -0.11917  0.82410  X74      -0.00594  0.51892
164 *X15      -0.62872 -0.45895  X15      -0.33918 -0.37232
165 *X25      -0.46845 -0.34602  X25      -0.32159 -0.35600
166 *X35      -0.34143 -0.24901  X35      -0.27432 -0.22658
167 *X45      -0.31265 -0.78521  X45      -0.30460 -0.40728
168 *X55      -0.25020 -0.66127  X55      -0.28930 -0.33741
169 *X65      -0.49599 -0.65055  X65      -0.30939 -0.38457
170 *X75      -0.70186  0.24134  X75      -0.47302  0.22995
171 *_
```

Jo näistä tuloksista on mahdollista päätellä jotain kummankin erottelumuuttujan luonteesta. Tulkinta helpottuu laskemalla erottelumuuttujien arvot ja tutkimalla niiden jakaumia eri tavoin.

```
 26  1 SURVO 84C EDITOR Sun Apr 24 17:13:01 1994        C:\M\MEN2\ 300 100 0
172 *
173 *Erottelumuuttujien D1,D2 laskeminen ja talletus:
174 *LINCO HIL,DISCRL.M(D1,D2)_
175 *
```

Huolimatta alkuperäisten 35 muuttujan dikotomisuudesta erottelumuuttujat näyttävät ryhmittäin tarkasteltuina likipitäen normaalisti jakautuneilta. Tässä on esimerkiksi otettu ensimmäisen erottelijan luokitettu jakauma ensimmäisessä ryhmässä (H) ja todettu sen yhteensopivuus normaalijakaumaan.

```
 20  1 SURVO 84C EDITOR Sun Apr 24 17:17:58 1994        C:\M\MEN2\ 300 100 0
175 *.....................................................................
176 *1. erottelumuuttujan normaalisuus 1. ryhmässä:
177 *GHISTO HIL,D1,CUR+1_ / D1=-8(0.5)0 FIT=NORMAL IND=K,1
178 *Frequency distribution of D1 in HIL: N=100
179 *
180 *Class midpoint  f     %   Sum     %     e      e    f     X^2
181 *     <=-7.00    0   0.0     0   0.0   1.1
182 *       -6.75    4   4.0     4   4.0   2.8
183 *       -6.25    8   8.0    12  12.0   6.7   10.6   12     0.2
184 *       -5.75   11  11.0    23  23.0  12.5   12.5   11     0.2
185 *       -5.25   18  18.0    41  41.0  18.1   18.1   18     0.0
186 *       -4.75   21  21.0    62  62.0  20.3   20.3   21     0.0
187 *       -4.25   16  16.0    78  78.0  17.5   17.5   16     0.1
188 *       -3.75   13  13.0    91  91.0  11.7   11.7   13     0.2
189 *       -3.25    5   5.0    96  96.0   6.0
190 *       -2.75    4   4.0   100 100.0   2.4
191 *      >-2.50    0   0.0   100 100.0   0.9    9.3    9     0.0
192 *Mean=-4.785000 Std.dev.=0.967613
193 *Fitted by NORMAL(-4.785,0.9363) distribution
194 *Chi-square=0.696 df=4 P=0.9518
195 *
```



/DISCRI-sukron muodostamien erottelumuuttujien korrelaatiomatriisien ja **MSN**-matriisien avulla on edullisinta luokitella sekä erotteluanalyysissa jo mukana olleita havaintoja sekä mahdollisia uusia havaintoja, jotka ovat aikaisemmista riippumattomia. On odotettavissa, että "vanhojen" havaintojen kohdalla luokittelutulos on liian optimistinen, koska ne itse vaikuttavat erottelijoihin.

Luokittelu tapahtuu uudella CLASSI-operaatiolla, joka käyttää korrelaatio- ja MSN-matriiseja ryhmien kuvaajina. Tämä komento on siis mahdollinen myös muille kuin erotteluanalyysilla muodostetuille matriiseille. Aluksi on määriteltävä ja valittava sopivat luokittelutuloksia kuvaavat muuttujat.

CLASSI sallii luokittelun yhtaikaa Mahalanobis-etäisyyksien ja Bayesin periaatteen mukaan. Edellisessä tapauksessa muuttuja aktivoidaan joko D- tai d-kirjaimella ja jälkimmäisessä tapauksessa B- tai b-kirjaimella. Isot kirjaimet tarkoittavat yhteisen kovarianssimatriisin käyttöä, pienet taas ryhmäkohtaisten. CLASSI tallettaa myös luokittelutodennäköisyydet (tai -etäisyydet tapauksissa D,d) P-kirjaimilla aktivoituihin muuttujiin, joita tulee olla sama määrä kuin vertailtavia ryhmiä. Jos käytetään yhtä useampaa luokittelusääntöä samanaikaisesti, nämä todennäköisyydet talletetaan sen kriteerin mukaan, joka on järjestyksessä b,B,d,D ensimmäisenä.

Tässä tapauksessa käytetään kaikkia luokittelusääntöjä rinnakkain. Todennäköisyydet tulevat olemaan siis Bayesin periaatteen mukaisia, kun ne lasketaan ryhmittäisten kovarianssien pohjalta.

```
 16  1 SURVO 84C EDITOR Sun Apr 24 19:51:53 1994        C:\M\MEN2\ 300 100 0
198 *Tulosmuuttujien määrittely havaintojen luokittelua varten:
199 *FILE UPDATE HIL_
200 *
201 *FIELDS:
202 *  39 ND-   1 Mahal1    Mahalanobis-etäisyys, samat kovarianssit
203 *  40 Nd-   1 Mahal2    Mahalanobis-etäisyys, ryhmittäiset kovarianssit
204 *  41 NB-   1 Bayes1    Bayes-todennäköisyys, samat kovarianssit
205 *  42 Nb-   1 Bayes2    Bayes-todennäköisyys, ryhmittäiset kovarianssit
206 *  43 NP-   4 PH        H-todennäköisyys
207 *  44 NP-   4 PI        I-todennäköisyys
208 *  45 NP-   4 PL        L-todennäköisyys
209 *END
210 *SURVO 84C data file HIL: record=128 bytes, M1=64 L=64  M=38 N=300
211 *
```

CLASSI edellyttää samanlaisia CORR- ja MSN-täsmennyksiä kuin /DISCRI-sukro. Nyt kannattaa viitata /DISCRI:n antamiin matriiseihin, jolloin luokittelu tapahtuu erottelumuuttujien avulla, vaikka niitä ei olisi valmiiksi laskettuna havainnoittain. Tässä tapauksessa CLASSI laskee erottelumuuttujien arvot luokittelun aikana kerroinmatriisin DISCRL.M avulla, johon viitataan rivin 216 COEFF-täsmennyksellä.

```
 11  1 SURVO 84C EDITOR Sun Apr 24 20:13:55 1994        C:\M\MEN2\ 300 100 0
212 *...........................................................................
213 *Havaintojen luokittelu erottelumuuttujien D1,D2 avulla:
214 *CORR=DR1,DR2,DR3
215 *MSN=DM1,DM2,DM3
216 *COEFF=DISCRL.M
217 *CLASSI HIL_
218 *
```

Eräs tapa selvittää erotteluanalyysin onnistuneisuutta on verrata alkuperäistä ryhmittelyä (muuttuja K) luokittelussa saatuihin tuloksiin (tässä esim. muuttujaan Bayes2). Se tapahtuu yksinkertaisesti taulukoimalla ko. muuttujat vastakkain:

```
 14  1 SURVO 84C EDITOR Mon Apr 25 10:38:09 1994        C:\M\MEN2\ 300 100 0
218 *.....................................................................
219 *Luokittelun tarkistus:
220 *TAB HIL,CUR+3_
221 *VARIABLES=Bayes2,K Bayes2=0,1(H),2(I),3(L) K=0,1(H),2(I),3(L)
222 *
223 *TABLE HIL1 A,B,F   N=300
224 A  Bayes2   H   I   L
225 *K ******
226 *H         95   0   5
227 *I          1  98   1
228 BL          6   3  91
229 *Chi_square=509.0 df=4 P=0.0000
230 *Virheluokituksia 100*16/300=5.3 %
231 *
```

Näemme, että 100 H-kirjaimesta 95 on luokiteltu oikein ja 5 on luokiteltu L-kirjaimiksi. Yhteensä 16 kirjainta 300:sta eli 5.3 % on luokiteltu väärin. Tässä tapauksessa kaikki neljä luokitustapaa johtavat täsmälleen samoihin tuloksiin, mikä ilmenee seuraavasta taulukoinnista:

```
 18  1 SURVO 84C EDITOR Mon Apr 25 11:30:31 1994        C:\M\MEN2\ 300 100 0
231 *.....................................................................
232 *Luokittelujen yhtäpitävyyden toteaminen:
233 *VARIABLES=Mahal1,Mahal2,Bayes1,Bayes2
234 *Mahal1=0,1(H),2(I),3(L)  Mahal2=0,1(H),2(I),3(L)
235 *Bayes1=0,1(H),2(I),3(L)  Bayes2=0,1(H),2(I),3(L)
236 *TAB HIL,CUR+2
237 *TABS HIL2,2,CUR+1_
238 *TABLE HIL2S E,F,F
239 E              Mahal1   H           I           L
240 *              Mahal2   H   I   L   H   I   L   H   I   L
241 *Bayes1 Bayes2 ******
242 *H      H             102   0   0   0   0   0   0   0   0
243 *       I               0   0   0   0   0   0   0   0   0
244 *       L               0   0   0   0   0   0   0   0   0
245 *I      H               0   0   0   0   0   0   0   0   0
246 *       I               0   0   0   0 101   0   0   0   0
247 *       L               0   0   0   0   0   0   0   0   0
248 *L      H               0   0   0   0   0   0   0   0   0
249 *       I               0   0   0   0   0   0   0   0   0
250 F       L               0   0   0   0   0   0   0   0  97
251 *
```

Kuten jo edellä todettiin, alkuperäisen aineiston luokittelu antaa liian myönteisen kuvan luokittelun onnistumisesta, koska luokittelukriteerit perustuvat samaan aineistoon. Tämän vuoksi olisi hyvä säästää, mikäli mahdollista, osa aineistosta pelkkää luokittelua varten.

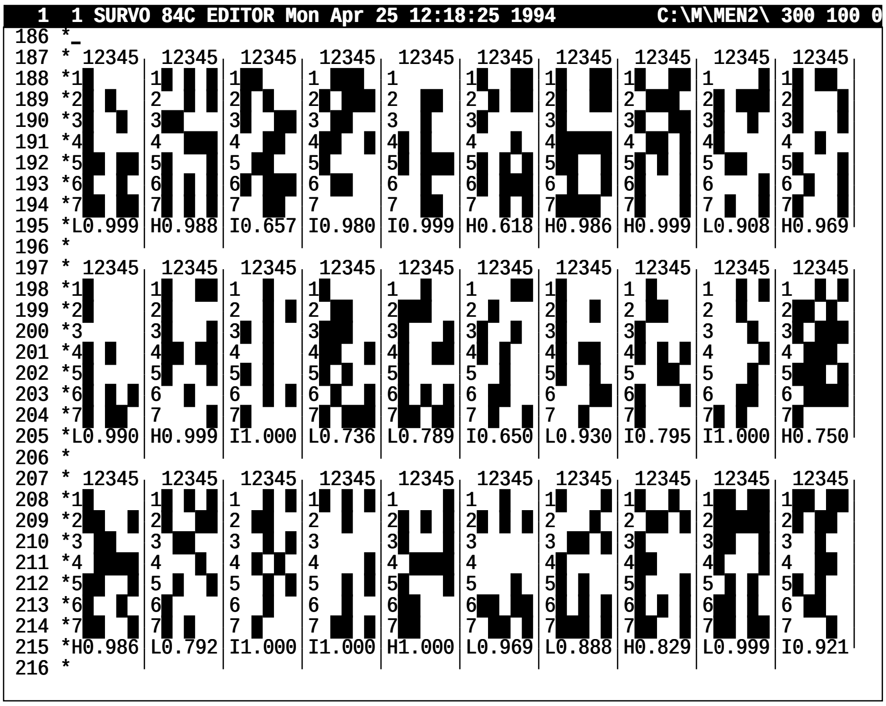
Tässä tapauksessa uusia, riippumattomia lisähavaintoja on tehtävissä vaikka kuinka paljon. Olemme luoneetkin toisen samanrakenteisen 3×100 havainnon aineiston `HIL2` käyttäen satunnaislukugeneraattorin `rand(100`1) asemasta generaattoria `rand(10011001)`. Tässä tapauksessa havaintojen luokittelu onnistuu seuraavasti:

```
 15  1 SURVO 84C EDITOR Mon Apr 25 12:03:24 1994         C:\M\MEN2\ 300 100 0
122 *...................................................................
123 *Rinnakkainen, riippumaton aineisto luotu generaattorilla rand(10011001)
124 *ja luokitettu samojen erottelumuuttujien mukaan:
125 *TAB HIL2,CUR+3_
126 *VARIABLES=Bayes2,K Bayes2=0,1(H),2(I),3(L) K=0,1(H),2(I),3(L)
127 *
128 *TABLE HIL21 A,B,F   N=300
129 A  Bayes2   H   I   L
130 *K ******
131 *H         88   2  10
132 *I          1  92   7
133 BL         10   4  86
134 *Chi_square=417.2 df=4 P=0.0000
135 *Virheluokituksia 100*34/300=11.3 %
136 *
```

Kuten oli odotettavissa, virheluokitusten määrä on suurempi ja tässä tapauksessa noin kaksinkertainen verrattuna alkuperäisen aineiston luokitteluun. Tätä 11.3 prosentin virheluokitusarviota on pidettävä uskottavana arviona erotteluanalyysin tarkkuudelle.

Aikaisempi 30 havainnon näyte (osa aineistoa `HIL2`), jonka annettiin lukijalle silmämääräisesti luokiteltavaksi on tässä esitetty uudelleen niin, että kunkin tapauksen alle on merkitty, mistä kirjaimesta on kysymys ja millä `Bayes2`-todennäköisyydellä se on tähän (oikeaan) ryhmään luokiteltu. Jokainen voi laskea oman virheluokitusprosenttinsa. Jos vääriä valintoja on alle 4 kappaletta, voi pitää itseään erotteluanalyysia parempana luokittelijana.

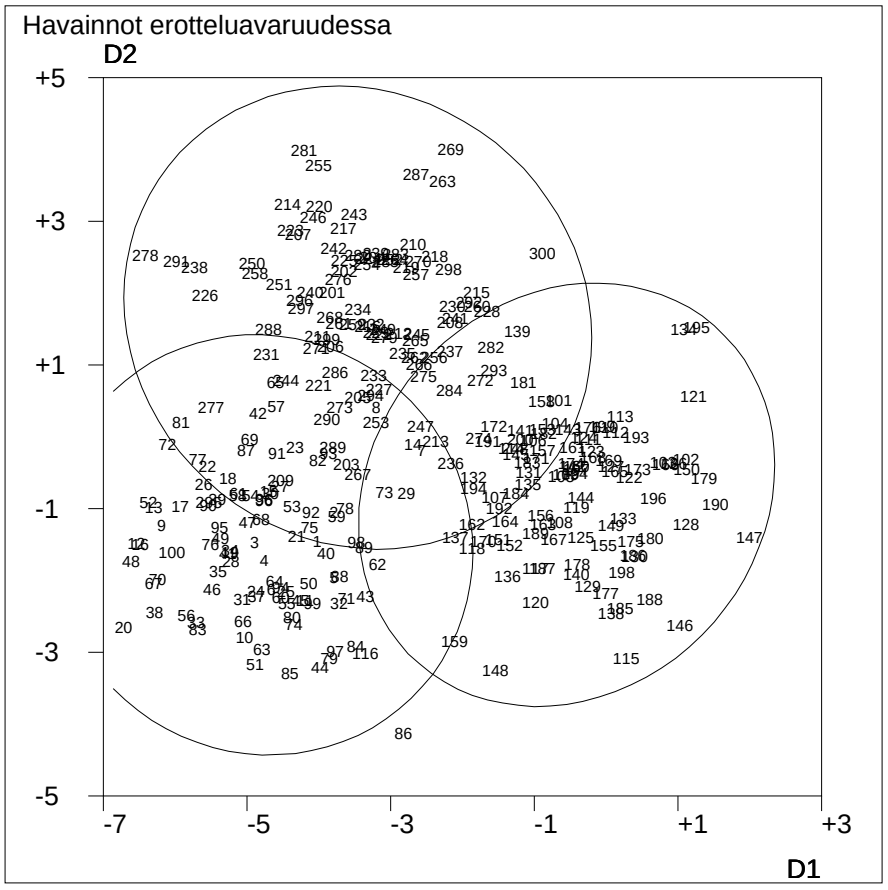
Nämä 30 tapausta on valittu siten, että useimmat ovat erotteluanalyysin mielestä täysin selviä. Kaikissa tapauksissa valintatodennäköisyys on ollut yli 0.6 . Näin silmämääräinen arviointi lienee helpompaa tästä näytteestä kuin se olisi koko aineistosta.



Kahden erottelumuuttujan kanssa on mukava piirtää kuvia erotteluavaruudesta. Esim. seuraava Survo-kaavio tekee kuvan, jossa havaintojen paikalla on niiden järjestysnumerot.

```
 16  1 SURVO 84C EDITOR Mon Apr 25 18:11:48 1994        C:\M\MEN2\ 100 100 0
  1 *
  2 *Ryhmien piirto erotteluavaruuteen (99%:n hajontaellipsit)
  3 *VAR NR:2=ORDER TO HIL
  4 *.....................................................................
  5 *HEADER=Havainnot_erotteluavaruudessa
  6 *GPLOT HIL,D1,D2  / MODE=VGA POINT=[SMALL],NR  IND=K,1
  7 *XSCALE=-7(2)3  YSCALE=-5(2)5  OUTFILE=A
  8 *CONTOUR=0.99
  9 *.....................................................................
 10 *GPLOT HIL,D1,D2  / MODE=VGA POINT=[RED][SMALL],NR  IND=K,2
 11 *XSCALE=-7(2)3  YSCALE=-5(2)5  OUTFILE=A INFILE=A
 12 *CONTOUR=0.99  HEADER=
 13 *.....................................................................
 14 *GPLOT HIL,D1,D2_ / MODE=VGA POINT=[GREEN][SMALL],NR  IND=K,3
 15 *XSCALE=-7(2)3  YSCALE=-5(2)5  OUTFILE=A INFILE=A
 16 *CONTOUR=0.99  HEADER=
 17 *.....................................................................
```

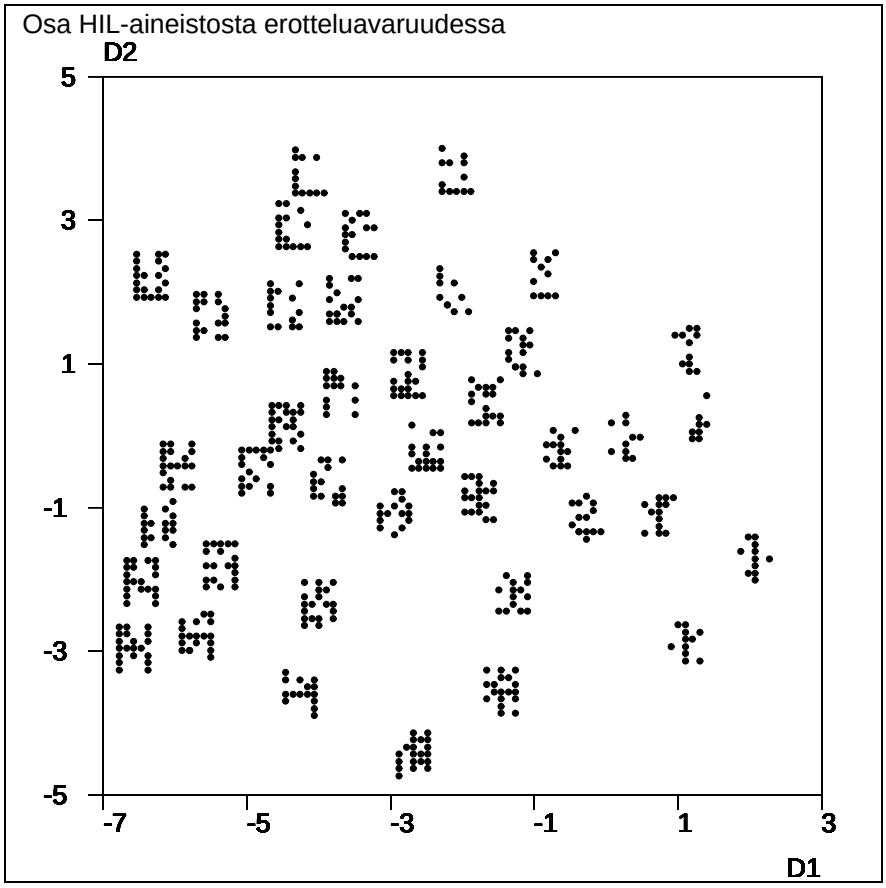
Tämä kuva paperille siirrettynä (ilman värejä) näyttäisi seuraavalta:



Kunkin ryhmän ympärillä on 99%:n hajontaellipsoidi. Erottelumuuttujat ovat miltei korreloimattomia ellipsien muodoista päätellen. Alla vasemmalla on H-kirjaimien ryhmä (numerot 1-100), siitä oikealle I-kirjaimet (numerot 101-200) ja ylinnä keskellä L-kirjaimet (numerot 201-300).

Näemme, että ensimmäinen erottelumuuttuja erottelee H- ja I-kirjaimet, toinen taas L-kirjaimet edellisistä.

Osa havainnoista on myös piirretty bittikarttoina tähän samaan erotteluavaruuteen:



Epämääräiset, vaikeasti luokiteltavat "kirjaimet" ovat kuvan keskellä. Puhtaimmat löytyvät kuvion reunoilta. Kunkin ryhmän osalta todennäköisimmät havainnot ovat tällaisia:

```
  1  1 SURVO 84C EDITOR Mon Apr 25 18:30:32 1994        C:\M\MEN2\ 100 100 0
 17 *.....................................................................
 18 *Parhaiten tunnistetut havainnot kussakin ryhmässä:
 19 *         20        147         281
 20 *       12345     12345       12345
 21 *     1 ▐▌ ▐▌   1 ▐▌ ▐▌     1 ▐▌
 22 *     2 ▐▌ ▐▌   2 ▐▌ ▐▌     2 ▐▌ ▐▌
 23 *     3 ▐▌ ▐▌   3 ▐▌ ▐▌     3
 24 *     4 ▐▌ ▐▌   4 ▐▌ ▐▌     4 ▐▌
 25 *     5 ▐▌ ▐▌   5 ▐▌ ▐▌     5 ▐▌
 26 *     6 ▐▌ ▐▌   6 ▐▌ ▐▌     6 ▐▌
 27 *     7 ▐▌ ▐▌   7 ▐▌ ▐▌     7 ▐▌▌▌
 28 *_
```

Vaikka esimerkkimme kertoo jotain hahmontunnistuksen ongelmista, se ei ole realistinen. Käytännössä eroteltavia kirjaimia tai muita objekteja saattaa olla kymmenittäin eikä vain kolme. Myös virhetodennäköisyyksien tulisi olla huomattavasti alle sen, mitä esimerkissä esiintyy. Toisaalta on huomattava, että "kohina" on esimerkissämme ollut tarkoituksellisesti poikkeuksellisen suurta niin, ettei silmämääräisesti liene mahdollista päästä samaan tarkkuuteen kuin erotteluanalyysilla. Siis huolimatta aineiston tilastollisista vajavuuksista erotteluanalyysi toimii hämmästyttävän hyvin.

Täydentäviä huomautuksia edelliseen esimerkkiin

Erottelutarkkuutta on mahdollista parantaa käyttämällä suurempaa perusaineistoa. Esimerkissämme näitä havaintoja oli 3×100. Nyt on tehty samat tarkastelut 3×5000 havainnon aineistolla käyttäen satunnaislukugeneraattoria `rand(3003)`. Tällöin luokittelu perusaineistosta antaa tuloksen

```
 15  1 SURVO 84C EDITOR Wed Apr 27 18:00:31 1994        C:\M\MEN2\ 400 100 0
221 *Luokittelun tarkistus:
222 *TAB HIL!,CUR+3_
223 *VARIABLES=Bayes2,K Bayes2=0,1(H),2(I),3(L) K=0,1(H),2(I),3(L)
224 *
225 *TABLE HIL!1 A,B,F   N=15000
226 A  Bayes2     H     I     L
227 *K ******
228 *H         4590    88   322
229 *I           86  4711   203
230 BL          372   214  4414
231 *Chi_square=22843 df=4 P=1.0000
232 *Virheluokituksia 100*1285/15000=8.57 %
233 *
```

eli virheluokituksia (8.57 %) on suhteellisesti *enemmän* kuin alkuperäisessä kokeessa, jossa niitä saatiin vain 5.3 % . Tämä johtuu siitä, että aikaisempi erotteluanalyysi taipui liikaa oman aineistonsa suuntaan, mitä ei pääse tapahtumaan samassa määrin suurilla otoksilla.

Niinpä käytettäessä edellisistä riippumatonta 3×1000 havainnon aineistoa (generaattori `rand(30033003)`), luokittelutulokseksi saadaan

```
 16  1 SURVO 84C EDITOR Wed Apr 27 18:10:14 1994        C:\M\MEN2\ 300 100 0
123 *Rinnakkainen, riippumaton aineisto luotu generaattorilla rand(30033003)
124 *ja luokitettu samojen erottelumuuttujien mukaan:
125 *TAB HIL!2,CUR+3_
126 *VARIABLES=Bayes2,K Bayes2=0,1(H),2(I),3(L) K=0,1(H),2(I),3(L)
127 *
128 *TABLE HIL!21 A,B,F   N=3000
129 A  Bayes2    H    I    L
130 *K ******
131 *H         912   18   70
132 *I          23  938   39
133 BL          59   56  885
134 *Chi_square=4524 df=4 P=1.0000
135 *Virheluokituksia 100*265/3000=8.83 %
136 *
```

eli virheluokitusten osuus (8.83 %) on noussut vain aavistuksen ja on pienempi kuin alkuperäisessä kokeessa riippumattomalla aineistolla saatu 11.3 %.

Tarkemmat erottelutulokset antavat myös luotettavammat arviot 30 havainnon näytteelle, jossa muutamien havaintojen luokittelutodennäköisyydet muuttuvat selvästi. Silmämääräisiä arvioita tulisi verrata tässä saatuihin luokituksiin.
Esim. kolmannen rivin toinen havainto osoittautuu hyvin epämääräiseksi, vaikka se alunperin on ollut L.

```
  1  1 SURVO 84C EDITOR Wed Apr 27 18:16:53 1994        C:\M\MEN2\ 400 100 0
255 *_
256 *  12345   12345   12345   12345   12345   12345   12345   12345   12345   12345
257 *1 ▄▄ ▄   1 ▄▄▄   1 ▄▄    1 ▄▄▄   1       1 ▄▄▄   1 ▄▄▄   1 ▄▄▄   1 ▄     1 ▄▄
258 *2 ▄  ▄   2 ▄  ▄   2 ▄ ▄   2 ▄▄    2 ▄▄    2 ▄▄    2 ▄▄    2 ▄▄    2 ▄▄    2 ▄  ▄
259 *3 ▄  ▄   3 ▄▄▄   3 ▄▄▄   3 ▄▄▄   3 ▄▄▄   3 ▄ ▄   3 ▄▄▄   3 ▄▄▄   3 ▄▄▄   3 ▄▄▄
260 *4 ▄  ▄   4 ▄▄▄   4 ▄▄▄   4 ▄ ▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄
261 *5 ▄▄ ▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄
262 *6 ▄▄ ▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄
263 *7 ▄▄ ▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄
264 *L1.000 H0.993 I0.924 I0.996 I1.000 H0.847 H0.986 H0.999 L0.865 H0.987
265 *                                   L0.144                H0.110
266 *
267 *  12345   12345   12345   12345   12345   12345   12345   12345   12345   12345
268 *1 ▄▄▄   1 ▄▄▄   1 ▄     1 ▄▄▄   1 ▄▄▄   1 ▄▄▄   1 ▄▄▄   1 ▄▄▄   1 ▄▄▄   1 ▄▄▄
269 *2 ▄     2 ▄▄▄   2 ▄  ▄   2 ▄▄▄   2 ▄▄▄   2 ▄▄▄   2 ▄     2 ▄     2 ▄▄▄   2 ▄▄▄
270 *3 ▄     3 ▄▄▄   3 ▄  ▄   3 ▄▄▄   3 ▄▄▄   3 ▄▄▄   3 ▄     3 ▄     3 ▄▄▄   3 ▄▄▄
271 *4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄
272 *5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄
273 *6 ▄▄▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄
274 *7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄
275 *L0.991 H0.997 I1.000 L0.969 L0.898 I0.653 L0.851 I0.736 I0.999 H0.926
276 *                           H0.101 L0.339 H0.132 L0.146       I0.074
277 *                                                 H0.118
278 *
279 *  12345   12345   12345   12345   12345   12345   12345   12345   12345   12345
280 *1 ▄▄▄   1 ▄▄▄   1 ▄▄▄   1 ▄▄▄   1 ▄▄▄   1 ▄▄▄   1 ▄▄▄   1 ▄▄▄   1 ▄▄▄   1 ▄▄▄
281 *2 ▄  ▄   2 ▄▄▄   2 ▄▄▄   2 ▄  ▄   2 ▄  ▄   2 ▄  ▄   2 ▄▄▄   2 ▄▄▄   2 ▄  ▄   2 ▄▄▄
282 *3 ▄▄▄   3 ▄▄▄   3 ▄▄▄   3 ▄  ▄   3 ▄▄▄   3 ▄▄▄   3 ▄▄▄   3 ▄▄▄   3 ▄▄▄   3 ▄▄▄
283 *4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄   4 ▄▄▄
284 *5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄   5 ▄▄▄
285 *6 ▄  ▄   6 ▄▄▄   6 ▄▄▄   6 ▄  ▄   6 ▄▄▄   6 ▄  ▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄   6 ▄▄▄
286 *7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄   7 ▄▄▄
287 *H0.973 I0.375 I1.000 I0.989 H1.000 I0.556 L0.986 L0.588 L0.773 I0.996
288 *       H0.342                  L0.443       H0.411 H0.227
289 *       L0.284
290 *
```

8. Ryhmittelyanalyysi

Ryhmittelyanalyysi (Cluster Analysis) kohdistuu tilastollisiin aineistoihin, jotka ovat useasta eri perusjoukosta saatujen otosten (ryhmien) yhdistelmiä. Tarkoituksena on paljastaa oikea ryhmien lukumäärä ja luokitella havainnot näihin ryhmiin. Ryhmittelyanalyysissa ei ole ryhmistä mitään ennakkotietoa kuten erotteluanalyysissa, joten tehtävä on hankalampi.

Koska pääkomponentti- ja faktorianalyysissa tavallaan ryhmitellään muuttujia, ryhmittelyanalyysi on jossain määrin rinnastettavissa myös näihin menetelmiin. Analyysin lähtökohtana on tällöin havaintomatriisin transpoosi tai jokin sen muunnos. Puhutaan esim. käännetystä faktorianalyysista.

Kaikki varsinaiset ryhmittelyanalyysin menetelmät käyttävät jonkinlaista ryhmien (ja yksittäisten havaintojen) välisen etäisyyden mittaa. Tarkoituksena on panna sellaiset havainnot yhteen, jotka ko. mitan suhteen ovat riittävän läheisiä.

Useimmat ryhmittelymenetelmät ovat luonteeltaan heuristisia; niiltä puuttuu selkeä teoreettinen tausta. Suosittuja ovat hierarkiset menetelmät. Tällöin esim. aluksi jokainen havainto muodostaa oman ryhmänsä ja etsitään ne havainnot, jotka ovat kaikkein läheisimpiä ja yhdistetään ne kahden havainnon ryhmäksi. Tämän jälkeen uusitaan sama menettely, jolloin syntyy toinen kahden havainnon ryhmä tai jokin havainto yhtyy ensimmäiseen kahden havainnon ryhmään. Tätä menettelyä toistetaan jatkuvasti, jolloin joka kerralla ryhmien lukumäärä vähenee yhdellä. Kun näin jatketaan, lopulta kaikki havainnot kasautuisivat yhdeksi ryhmäksi. Tarkoitus on kuitenkin keskeyttää menettely sellaiseen vaiheeseen, jossa esim. ryhmittelyn hyvyyttä kuvaavassa kriteerissä tapahtuu selvä muutos. Hierarkinen ryhmittely voi tapahtua myös toisinpäin lähtemällä jakamaan kaikkien havaintojen muodostamaa ryhmää vaiheittain pienempiin.

Kokonaan toisenlainen on menettely, jossa etukäteen tiedetään tai arvioidaan oikea ryhmien lukumäärä g ja jaetaan havainnot aluksi umpimähkään g ryhmään. Käyttämällä mittaa, joka kuvaa ryhmien homogeenisuutta, pyritään parantamaan tilannetta havaintojen siirroilla ryhmästä toiseen. Yksinkertaisimmillaan menettely on sellainen, että käydään havaintoja systemaattisesti läpi ja koemielessä yritetään siirtää tarkasteltavaa havaintoa toisiin ryhmiin. Heti kun siirto parantaa ryhmien homogeenisuutta, annetaan sen toteutua ja jatketaan menettelyä seuraavasta havainnosta. Lopullinen ratkaisu saavutetaan, kun minkään havainnon siirto ei enää paranna ryhmien homogeenisuutta. Tämä ei takaa välttämättä, että saavutettaisiin paras ratkaisu, koska satunnainen alkuryhmitys usein vaikuttaa siihen, mihin lopulta päädytään. Hyvän ratkaisun löytämiseksi koko menettely on tällöin syytä toistaa riittävän monta kertaa lähtien aina uudesta satunnaisryhmityksestä ja valita tuloksista paras.

8.1 Tilastollinen ryhmittelyanalyysi

Tässä yhteydessä kuvaamme vain ns. tilastollista ryhmittelyanalyysia, jossa noudatetaan viimeksi mainittua ryhmittelymenettelyä käyttäen ryhmien homogeenisuuden mittana ns. *Wilks*in lambda-kriteeriä *L*. Erotteluanalyysissa käytettyjen merkintöjen mukaan on

$$L = |\mathbf{W}|/|\mathbf{T}|$$

ja yhtälöiden **A'WA=I** ja **A'TA=Λ+I** perusteella on

$$1/L = \prod_{i=1}^{p}(1 + \lambda_i)$$

eli log *L* on vakiotekijää vaille erotteluanalyysissa esiintyneen ryhmien keskiarvovektoreiden samuutta kuvaava testisuure. *L* mittaa ryhmien homogeenisuutta siten, että se tulee sitä pienemmäksi mitä homogeenisempia ryhmät ovat ja mitä paremmin ne erottuvat toisistaan.

Multinormaalisissa ryhmissä, joilla on samat kovarianssit, Wilksin lambda-kriteeri on ilmeisesti paras mahdollinen. Sen käytön esteenä olivat pitkään laskutekniset hankaluudet, koska jokaisen havainnon siirtokokeilun yhteydessä matriisi **W** ja sen determinantti muuttuu. *Pekka Korhonen* onnistui vuonna 1977 kehittämään *Hannu Väliahon* regressioanalyysiin liittyvien, askeltavia algoritmeja koskevien tutkimusten pohjalta vaiheittaisen menettelyn, jolla ko. determinantin arvo voidaan päivittää mahdollisimman vähin laskutoimituksin havainnon siirtyessä ryhmästä toiseen. Aikaisemmin oli tyydytty helpommin laskettavaan mutta periaatteessa huonompaan standardoituun minimivarianssikriteeriin $\mathrm{tr}(\mathbf{WT}^{-1})$.

Ennen ryhmittelyä kannattaa yleensä vähentää muuttujien lukumäärää esim. pääkomponentti- tai faktorianalyysin avulla samoista syistä, jotka mainittiin erotteluanalyysin luokittelutehtävän yhteydessä. Tämä myös nopeuttaa ryhmittelyanalyysin suoritusta ja antaa näin paremmat mahdollisuudet etsiä ratkaisua usealla alkuryhmityksellä.

Jos oikeata ryhmien määrää ei etukäteen tiedetä, joudutaan lisäksi kokeilemaan eri *g*:n arvoja. Oikea ryhmien määrä päätellään lambda-arvojen perusteella.

Survossa ryhmittelyanalyysi tehdään Wilksin lambda-kriteeriä käyttäen `CLUSTER`-operaatiolla. Ryhmien määrä osoitetaan `GROUPS`-täsmennyksellä. Oletusarvo on `GROUPS=2`. Ryhmittelyn perustana olevat muuttujat aktivoidaan A-kirjaimilla ja ryhmittelyn tulos G-kirjaimella. Muodostuneet ryhmät osoitetaan indeksein 1,2,...,g, jotka tallentuvat G-muuttujaan. Ryhmittely voidaan toistaa erilaisista satunnaisista alkuryhmityksistä lähtien. Toistokertojen lukumäärä annetaan `TRIALS`-täsmennyksellä. Tällöin G-muuttujia voi nimetä useita, jolloin `CLUSTER` pitää kirjaa yhtä monesta parhaasta ratkaisusta, tallettaa

niitä vastaavat ryhmitykset ja tulostaa eri ratkaisuja vastaavat lambda-arvot sekä niiden frekvenssit.

Alkuryhmityksien arvonnassa käytettävä satunnaislukugeneraattori nimetään RND- (tai SEED-) täsmennyksellä. Alkuryhmityksen voi osoittaa myös I-kirjaimella aktivoidulla muuttujalla, jolloin tämän muuttujan arvojen tulee olla 1,2,...,g . Muut arvot tulkitaan puuttuviksi. Jos alkuryhmityksessä on puuttuvia arvoja, ao. havainnot luokitellaan alustavasti liittämällä ne lähimpään ryhmään ortogonalisoidussa havaintomatriisissa euklidisen etäisyysmitan mukaan. Näin soveltaja voi tarvittaessa ohjata ryhmittelyä valitsemiensa avainhavaintojen ympärille.

8.1.1 Esimerkki 1

Kokeilemme aluksi ryhmittelyanalyysia erotteluanalyysin yhteydessä kuvattuun HIL-aineistoon. Tässä tapauksessa tiedämme jokaisen havainnon alkuperäisen ryhmän. Tarkoituksena on vain katsoa, kuinka hyvin ryhmittelyanalyysi löytää ne.

Ei tietenkään ole järkeä käyttää alkuperäisiä dikotomisia muuttujia, joita on 35, sellaisenaan ryhmittelyyn, vaan teemme ensin faktorianalyysin ja laskemme faktoripistemäärät, jotka sitten otamme ryhmittelyn pohjaksi. Osoittautuu, että kunnollisia faktoreita löytyy vain kaksi. Kolmannen mukaanotto ei parantaisi ryhmittelytulosta.

Seuraavassa kaaviossa on esitetty faktorianalyysin eri vaiheet Varimax-ratkaisua myöten.

```
 22  1 SURVO 84C EDITOR Sat Apr 30 17:56:16 1994        C:\M\MEN2\ 100 100 0
  1 *
  2 *MASK=--AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
  3 *CORR HIL
  4 *FACTA CORR.M,2
  5 *ROTATE FACT.M,2,CUR+1_
  6 *Rotated factor matrix AFACT.M=FACT.M*TFACT.M
  7 *             F1     F2 Sumsqr
  8 *X11       0.390 -0.093  0.161
  9 *X21       0.261 -0.015  0.069
 10 *X31       0.386 -0.002  0.149
 11 *X41       0.524  0.037  0.276
 12 *X51       0.392  0.009  0.154
 13 *X61       0.340 -0.006  0.116
 14 *X71       0.303 -0.062  0.096
 15 *X12       0.065  0.065  0.008
 16 *X22       0.032 -0.096  0.010
 17 *X32      -0.076  0.033  0.007
 18 *X42       0.305  0.326  0.199
 19 *X52       0.026 -0.026  0.001
 20 *X62       0.111 -0.020  0.013
 21 *X72       0.157 -0.267  0.096
 22 *X13      -0.348  0.090  0.129
 23 *X23      -0.395  0.248  0.218
 24 *X33      -0.304  0.056  0.096
 25 *X43      -0.060  0.317  0.104
 26 *X53      -0.376  0.118  0.156
 27 *X63      -0.410  0.104  0.179
 28 *X73      -0.161 -0.291  0.111
 29 *X14       0.090  0.022  0.009
 30 *X24       0.008 -0.023  0.001
 31 *X34       0.004  0.009  0.000
 32 *X44       0.255  0.334  0.177
 33 *X54       0.107 -0.018  0.012
 34 *X64       0.045 -0.022  0.003
 35 *X74       0.066 -0.493  0.247
 36 *X15       0.205  0.403  0.204
 37 *X25       0.244  0.341  0.176
 38 *X35       0.197  0.207  0.082
 39 *X45       0.269  0.267  0.143
 40 *X55       0.239  0.272  0.131
 41 *X65       0.200  0.297  0.128
 42 *X75       0.411 -0.167  0.197
 43 *Sumsqr    2.426  1.429  3.855
```

Lasketaan sitten faktoripistemäärät (F1,F2) ja määritellään tiedostoon HIL kolme uutta muuttujaa (G1,G2,G3) ryhmittelytuloksia varten.

```
 16  1 SURVO 84C EDITOR Sat Apr 30 18:05:22 1994        C:\M\MEN2\ 100 100 0
 47 *
 48 */FCOEFF AFACT.M
 49 *Use FCOEFF.M for factor scores by LINCO <data>,FCOEFF.M(F1,F2,...)
 50 *LINCO HIL,FCOEFF.M(F1,F2)
 51 *
 52 *FILE UPDATE HIL_
 53 *
 54 *FIELDS: (active)
 55 *  49 NAG   1 G1
 56 *  50 NAG   1 G2
 57 *  51 NAG   1 G3
 58 *END
```

Ryhmittelyanalyysi tapahtuu seuraavasti:

```
 18  1 SURVO 84C EDITOR Sat Apr 30 18:09:27 1994        C:\M\MEN2\ 100 100 0
 59 *.....................................................................
 60 *VARS=F1(A),F2(A),G1(G),G2(G),G3(G)
 61 *CLUSTER HIL,CUR+2_ / TRIALS=10 RND=rand(19941994) GROUPS=3
 62 *
 63 *Stepwise cluster analysis by Wilks' Lambda criterion
 64 *Data HIL  N=300
 65 *Variables: F1, F2
 66 *Best clusterings found in 10 trials are saved as follows:
 67 * Lambda     freq  Grouping var
 68 * 0.08614      9   G1
 69 * 0.08615      1   G2
 70 *
```

Rivillä 60 on kuvattu muuttujien valinta ja niiden tehtävät analyysissa. Komentorivillä 61 on lisätäsmennyksenä koetoistojen määrä 10, satunnaislukugeneraattori ja ryhmien lukumäärä 3.

Kymmenessä yrityksessä on löytynyt vain kahdenlaisia ratkaisuja, joista lievästi parempi on tullut esiin 9 kertaa. Tämä viittaa siihen, ettei tulos tästä parane, vaikka koe uusittaisiin useampiakin kertoja.

Seuraavat taulukoinnit osoittavat, kuinka hyvin ryhmittelyanalyysi on tunnistanut alkuperäisen ryhmäjaon:

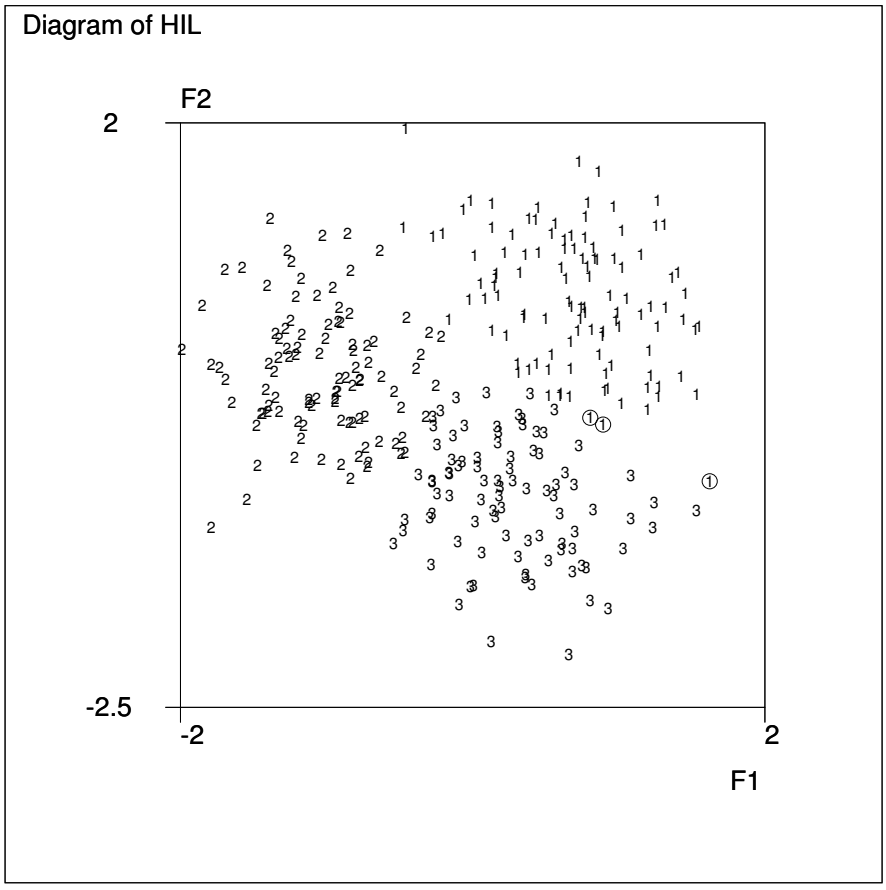
```
  1  1 SURVO 84C EDITOR Sat Apr 30 19:54:02 1994        C:\M\MEN2\ 120 100 0
 71 *.....................................................................
 72 *TAB HIL,CUR+2
 73 *VARIABLES=G1,K K=0,1(H),2(I),3(L) G1=0,1,2,3
 74 *TABLE HIL1 A,B,F   N=300
 75 A  G1   1   2   3
 76 *K **
 77 *H     95   1   4
 78 *I      4  94   2
 79 BL     12   5  83
 80 *Chi_square=447.0 df=4 P=0.0000
 81 *Virheluokituksia 100*28/300=9.3 %
 82 *.....................................................................
 83 *TAB HIL,CUR+2
 84 *VARIABLES=G2,K K=0,1(H),2(I),3(L) G2=0,1,2,3
 85 *TABLE HIL1 C,D,F   N=300
 86 C  G2   1   2   3
 87 *K **
 88 *H      4   1  95            G2-ryhmityksessä ryhmien
 89 *I      2  94   4            numerointi "sattumalta" päinvastainen
 90 DL     86   5   9
 91 *Chi_square=461.1 df=4 P=0.0000
 92 *Virheluokituksia 100*25/300=8.3 %
 93 *_
```

On aika hämmästyttävää, että tässä sovelluksessa ryhmittely onnistuu suurin piirtein yhtä hyvin kuin erotteluanalyysin jälkeinen luokittelu. Osittain tämä johtuu siitä, että aineisto on suhteellisen siisti ja ryhmät erottuvat melko hyvin toisistaan.

Lambda-kriteerin mielessä molemmat ratkaisut ovat yhtä hyviä. Sama koskee luokittelutarkkuutta. Taulukoimalla ryhmittelyt `G1` ja `G2` vastakkain nähdään, miten ne eroavat toisistaan:

```
 14  1 SURVO 84C EDITOR Sat Apr 30 20:04:04 1994        C:\M\MEN2\ 120 100 0
 93 *.....................................................................
 94 *TAB HIL,CUR+2_
 95 *VARIABLES=G2,G1 G2=0,1,2,3 G1=0,1(H),2(I),3(L)
 96 *TABLE HIL1 E,F,F   N=300
 97 E   G2   1   2   3
 98 *G1 **
 99 *H       3   0 108
100 *I       0 100   0
101 FL      89   0   0
102 *Chi_square=582.4 df=4 P=0.0000
```

Näemme, että ryhmittelyt ovat lähes identtiset. G1-ryhmittelyn 3 H-kirjainta on tulkittu G2-ryhmittelyssä L-kirjaimiksi ja ilmeisesti G2-ryhmittely on tässä suhteessa oikeassa. Nämä erimielisyyttä aiheuttavat havainnot näkyvät seuraavassa kuvassa ympyröityinä.



8.1.2 Esimerkki 2

Tilastollisessa ryhmittelyanalyysissa voi saada järjettömiäkin tuloksia, ellei pidä huolta siitä, että ryhmittely toistetaan riittävän monta kertaa eri alkuryhmityksistä lähtien.

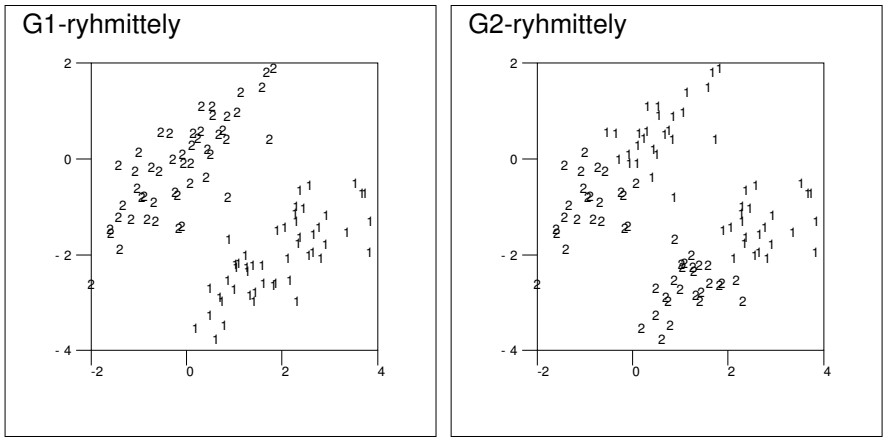
```
 25  1 SURVO 84C EDITOR Sun May 01 10:06:14 1994          C:\M\MEN2\ 100 100 0
  1 *
  2 *FILE CREATE N2,32,10,64,7,100
  3 *FIELDS:
  4 *1 N 4 X
  5 *2 N 4 Y
  6 *END
  7 *
  8 *VAR X,Y TO N2
  9 *X=if(ORDER<51)then(X1)else(X2) Y=if(ORDER<51)then(Y1)else(Y2)
 10 *X1=Z1     Y1=r*Z1+s*Z2      r=0.8 s=sqrt(1-r*r)
 11 *X2=Z1+2   Y2=r*Z1+s*Z2-2
 12 *Z1=probit(rand(1051994)) Z2=probit(rnd(1051994))
 13 *.....................................................................
 14 *VAR G1:1,G2:1,G3:1 TO N2_
 15 *G1=0 G2=0 G3=0
 16 *
```

Tässä on luotu kaksi 50 havainnon otosta 2-ulotteista normaalijakaumista, joilla on sama korrelaatiokerroin 0.8 ja samat hajonnat (1,1) mutta eri odotusarvovektorit eli ensimmäisessä ryhmässä (0,0) ja jälkimmäisessä (2,-2). Erillisellä VAR-operaatiolla on määritelty kolme 1-tavuista muuttujaa (G1,G2,G3) ryhmittelytuloksia varten.

```
 17  1 SURVO 84C EDITOR Sun May 01 10:16:18 1994          C:\M\MEN2\ 100 100 0
 16 *.....................................................................
 17 *MASK=AAGGG    TRIALS=10  GROUPS=2 RND=rand(51994)
 18 *CLUSTER N2,CUR+1_
 19 *Stepwise cluster analysis by Wilks' Lambda criterion
 20 *Data N2  N=100
 21 *Variables: X, Y
 22 *Best clusterings found in 10 trials are saved as follows:
 23 * Lambda     freq  Grouping var
 24 * 0.04715      5   G1
 25 * 0.14904      5   G2
 26 *
```

Tilastollinen ryhmittelyanalyysi on antanut kaksi erilaista ratkaisua, kumpaakin tässä tapauksessa 5 kertaa. Lambda-kriteerin mukaan ensimmäinen on selvästi parempi. Jos esitetään kummankin ryhmittelyn tulokset kuvallisesti esim. GPLOT-kaaviolla,

```
 13  1 SURVO 84C EDITOR Sun May 01 10:21:35 1994          C:\M\MEN2\ 100 100 0
 27 *.....................................................................
 28 *GPLOT N2,X,Y_ / HEADER=G1-ryhmittely
 29 *POINT=G1  (G2 antaa huonon luokittelun)
 30 *
```



`G1`-ryhmittely antaa odotetusti oikean tuloksen. `G2`-ryhmittely "näkee" muuttujien riippuvuuden toisinpäin. Näin tapahtuu, jos satunnaisessa alkuryhmityksessä havainnot asettuvat tällaista näkemystä suosivaksi. Askeltavalle havaintojen siirrolle ei jää mitään mahdollisuuksia irtaantua tästä virheellisestä asetelmasta.

Tämän vuoksi on tärkeätä, että ryhmittely toistetaan riittävän monta kertaa, jottei syntyisi pelkästään epäolennaisia ratkaisuja.

9. Moniulotteinen skaalaus

Moniulotteisella skaalauksella (multidimensional scaling) tarkoitetaan menettelyä, jolla *n* havainnon välisiä eroja koskevien etäisyys-, samanlaisuus- tai erilaisuustietojen perusteella havaintoja vastaavat pisteet yritetään sijoittaa "kartalle" eli tavallisesti 1- , 2- tai korkeintaan 3-ulotteiseen avaruuteen siten, että näiden pisteiden keskinäiset etäisyydet vastaavat annettuja etäisyystietoja.

Tavallisesti lähtökohtana on $n \times n$-etäisyys- tai erilaisuusmatriisi **D**, joka on ontto siten, että lävistäjäalkiot ovat nollia (eli havainnon etäisyys itseensä on 0). Muut alkiot ovat ei-negatiivisia. Yleensä **D** on symmetrinen, mutta esim. epälineaarisessa skaalauksessa saatetaan käsitellä myös epäsymmetrisiä tilanteita. Samoin etäisyystiedot voivat olla puuttuvia tiettyyn rajaan asti.

Jos lähtötiedot koskevat vertailtavien havaintojen samanlaisuutta, ne tulee muuntaa erilaisuutta kuvaaviksi esim. vähentämällä ne vakiosta tai siirtymällä käänteisarvoihin jne. Yleensäkin tarvittaessa etäisyysmatriisi olisi hyvä ensin transformoida alkioittain sellaiseen muotoon, että on syytä uskoa arvojen kuvaavan havaintopisteiden välisiä euklidisia etäisyyksiä. $n \times n$-neliömatriisia, jonka alkiot on tulkittavissa *n* pisteen välisiksi euklidisiksi etäisyyksiksi *p*-ulotteisessa ($p<n$) avaruudessa, sanomme euklidiseksi matriisiksi.

Jos moniulotteinen skaalaus halutaan tehdä havaintoaineistosta, jossa on useita asiaan vaikuttavia muuttujia, etäisyysmatriisi **D** tulee ensin laskea näiden muuttujien perusteella esim. havaintojen euklidisina tai Mahalanobis-etäisyyksinä. Tätä varten Survossa on erityinen `DIST`-operaatio, joka muodostaa etäisyysmatriiseja lukuisien erilaisten etäisyysmittojen mukaan. Mitta-arvoja laskettaessa muuttujat voidaan sekä standardoida että painottaa. Etäisyysmitat on esitelty tarkemmin esim. teoksessa *Cox-Cox* (1994).

Joskus halutaan tutkia myös muuttujien välisiä "etäisyyksiä". Silloin on etäisyysmatriisien laskentaan käytettävissä Survon `DISTV`-operaatio.

Moniulotteinen skaalaus liittyy monimuuttujamenetelmiin erityisesti siten, että ns. klassisessa skaalauksessa ongelma palautuu etäisyysmatriisin tietyllä muunnoksella pääkomponenttianalyysiin. Klassista skaalausta on ensimmäisenä käsitellyt *Richardson* v.1938 ja se on tullut yleisemmin tunnetuksi *Torgerson*in tutkimusten (1952) kautta.

Klassinen skaalaus ei kuitenkaan aina toimi tyydyttävästi etenkään, jos etäisyysmatriisia ei saada muokatuksi euklidiseksi. Monissa sovelluksissa havaintojen etäisyyksien arvioiminen on siinä määrin vaikeaa, että mittaus on vain järjestysasteikollista. Tähän perustuu mm. *Shepard*in ja *Kruskal*in 1970-luvulla kehittämä ordinaaliskaalaus, jossa etäisyyksille etsitään iteratiivisesti paras aineistokohtainen monotoninen muunnos niin, että havaintojen esittäminen vähäulotteisessa avaruudessa onnistuu.

Voidaan tietenkin kysyä, miksi edes etäisyyksien järjestyksestä tulisi välttämättä pitää kiinni. Kokemusten mukaan (*Chatfield*, *Collins* s. 210) ainakin yhtä hyvin kuin ordinaaliskaalaus toimii suora pienimmän neliösumman keino, jossa annetulle etäisyysmatriisille etsitään havaintojen koordinaattiesitys niin, että havaittujen ja koordinaattiesityksestä laskettujen etäisyyksien (mahdollisesti painotettu) neliösumma minimoituu. Tämä menettely on periaatteessa yksinkertaisin mutta laskennallisesti raskain.

9.1 Klassinen skaalaus

Tarkastelemme asiaa aluksi toisin päin olettaen, että n pisteen koordinaatit p-ulotteisessa avaruudessa muodostavat $n\times p$-matriisin (konfiguraation) **X**. Käytämme pisteiden r ja s (r,s=1,2,...,n) väliselle etäisyydelle merkintää d_{rs} . Tällöin tämän etäisyyden neliö voidaan lausua matriisin **B=XX**' alkioiden

$$b_{rs} = \sum_{j=1}^{p} x_{rj}x_{sj}$$

avulla muodossa

$$(1)\quad d_{rs}^2 = \sum_{j=1}^{p} (x_{rj} - x_{sj})^2 = b_{rr} + b_{ss} - 2b_{rs} .$$

On selvää, että jos tunnetaan etäisyydet **D**=[d_{rs}], matriisi **X** ei määräydy yksikäsitteisesti. On esim. lupa siirtää koordinaatistoa siten, että havaintopisteiden keskiarvo osuu origoon eli

$$\sum_{r=1}^{n} x_{rj} = 0 , \; j = 1,2,..., p .$$

Tästä seuraa, että matriisin **B** sekä vaaka- että pystyrivisummat ovat nollia eli

$$\sum_{r=1}^{n} b_{rs} = 0 , \; s = 1,2,..., n , \quad \sum_{s=1}^{n} b_{rs} = 0 , \; r = 1,2,..., n .$$

Näiden ja yhtälön (1) nojalla saamme

$$\sum_{r=1}^{n} d_{rs}^2 = T + n b_{ss} , \quad \sum_{s=1}^{n} d_{rs}^2 = n b_{rr} + T$$

ja

$$\sum_{r=1}^{n} \sum_{s=1}^{n} d_{rs}^2 = 2nT ,$$

missä T=tr(**B**) .

Edelleen yhtälön (1) ja kolmen viimeisen yhtälön perusteella saadaan

$$-2b_{rs} = d_{rs}^2 - d_{r.}^2 - d_{.s}^2 + d_{..}^2 ,$$

missä $d_{r.}^2$, $d_{.s}^2$ ja $d_{..}^2$ tarkoittavat r. rivin, s. sarakkeen ja kokonaiskeskiarvoa

etäisyyksien d_{rs} neliöistä.

Matriisi **B** syntyy näin ollen etäisyysmatriisista **D** korottamalla **D**:n alkiot toiseen potenssiin, jakamalla saatu matriisi luvulla -2 ja lopuksi keskistämällä se ensin rivien suhteen ja sitten sarakkeiden suhteen.

On ilmeistä, että välttämätön ja riittävä ehto sille, että etäisyysmatriisi **D** on euklidinen, on se, että **B** on ei-negatiivisesti definiitti, koska **B=XX'** .

Konfiguraatiomatriisi **X** ei määräydy **D**:n (ja siis myös **B**:n) avulla yksikäsitteisesti. Paitsi keskistystä (**X**:n sarakesummat = 0) tulokseen jää faktorianalyysin tapaan (ortogonaalinen) rotaatiomahdollisuus. Jos **T** on $p \times p$-ortogonaalinen matriisi, myös **Y=XT** kelpaa ratkaisuksi, sillä **YY'=XTT'X'=XX'=B** .

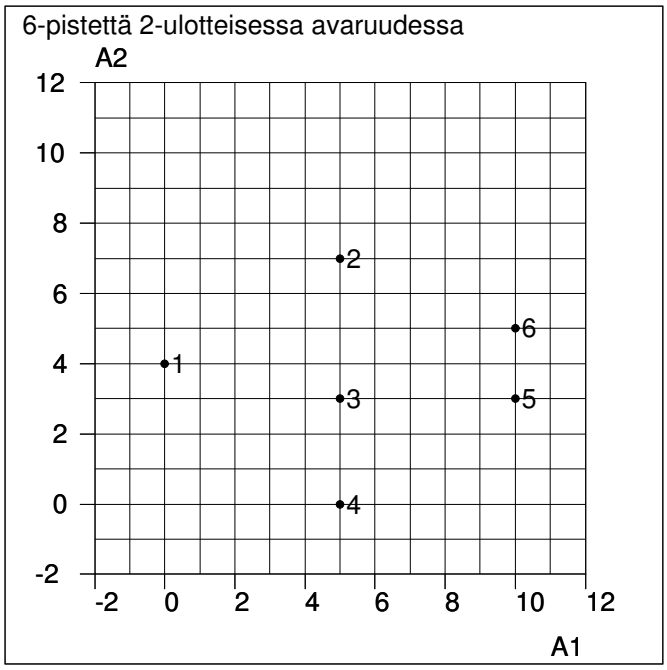
Jos rinnastamme matriisin **B** kovarianssimatriisiin, ratkaisu **X** voidaan valita siten, että $\mathbf{X}=P^{(p)}(\mathbf{B})$ eli **B**:n p ensimmäistä pääkomponenttia antaa tuloksen.

9.1.1 Esimerkki 1

Survossa klassinen skaalaus tehdään sukrolla `/CSCAL`. Näytämme kuitenkin ensin (ilman sukroa) vaiheittain matriisikäskyillä, miten tämä skaalaus toimii käytännössä. Tarkastelemme puhdasta kaksiulotteista, 6 pisteen muodostamaa tilannetta,

```
 11  1 SURVO 84C EDITOR Wed Jun 08 08:23:30 1994        D:\M\MEN2\ 240 100 0
  1 *
  2 *MATRIX X
  3 *///  A1 A2
  4 *1     0  4
  5 *2     5  7
  6 *3     5  3
  7 *4     5  0
  8 *5    10  3
  9 *6    10  5
 10 *
 11 *MAT SAVE X_
 12 *
```

joka kuviona näyttää seuraavalta:



Kuvasta on helppo laskea pisteiden välisten etäisyyksien neliöt (Pythagoraan lauseella). Talletetaan nämä matriisina `D2` . Sama tehdään myös automaattisesti `DIST`-operaatiolla, joka ilman lisätäsmennyksiä antaa itse etäisyydet.

```
 12  1 SURVO 84C EDITOR Wed Jun 08 08:28:32 1994          D:\M\MEN2\ 240 100 0
 12 *
 13 *MATRIX D2
 14 *// X1  X2  X3  X4  X5  X6
 15 *X1   0  34  26  41 101 101
 16 *X2  34   0  16  49  41  29
 17 *X3  26  16   0   9  25  29
 18 *X4  41  49   9   0  34  50
 19 *X5 101  41  25  34   0   4
 20 *X6 101  29  29  50   4   0
 21 *
 22 *MAT SAVE D2_
 23 *
```

Yritämme nyt tämän etäisyyksien neliöiden matriisin pohjalta määrätä konfiguraatiomatriisin **X**.

Tätä varten määräämme ensin matriisin **B** kertomalla matriisin `D2` luvulla -1/2 ja keskistämällä sekä pysty- että vaakarivien suhteen. Se tapahtuu Survossa yksinkertaisesti matriisioperaatiolla `CENTER`, sitten transponoimalla ja lopuksi uudelleen `CENTER`-operaatiolla:

```
 17  1 SURVO 84C EDITOR Wed Jun 08 08:40:35 1994        D:\M\MEN2\ 240 100 0
 23 *
 24 *MAT A=-0.5*D2      / *A~(-0.5)*D2 S6*6
 25 *MAT B=CENTER(A)    / *B~CENTER((-0.5)*D2) 6*6
 26 *MAT B=B'           / *B~CENTER((-0.5)*D2)' 6*6
 27 *MAT B=CENTER(B)    / *B~CENTER(CENTER((-0.5)*D2)') 6*6
 28 *
 29 *MAT LOAD B,CUR+1_
 30 *MATRIX B
 31 *CENTER(CENTER((-0.5)*D2)')
 32 *///            X1       X2       X3       X4       X5       X6
 33 *X1         34.1389   5.9722   4.6389   3.6389 -24.5278 -23.8611
 34 *X2          5.9722  11.8056  -1.5278 -11.5278  -5.6944   0.9722
 35 *X3          4.6389  -1.5278   1.1389   3.1389  -3.0278  -4.3611
 36 *X4          3.6389 -11.5278   3.1389  14.1389  -1.0278  -8.3611
 37 *X5        -24.5278  -5.6944  -3.0278  -1.0278  17.8056  16.4722
 38 *X6        -23.8611   0.9722  -4.3611  -8.3611  16.4722  19.1389
 39 *
```

Teemme nyt **B**:lle "pääkomponenttianalyysin" laskemalla ominaisarvot L ja -vektorit S.

```
 36  1 SURVO 84C EDITOR Wed Jun 08 08:47:35 1994        D:\M\MEN2\ 240 100 0
 39 *
 40 *MAT SPECTRAL DECOMPOSITION OF B TO S,L
 41 *MAT LOAD L,##.################,CUR+1_
 42 *MATRIX L
 43 *L(CENTER(CENTER((-0.5)*D2)'))
 44 *///                eigenval
 45 *ev1      70.897096827461920
 46 *ev2      27.269569839204750
 47 *ev3      -0.000000000000000
 48 *ev4      -0.000000000000001
 49 *ev5      -0.000000000000003
 50 *ev6      -0.000000000000006
 51 *
```

Kuten oli odotettavissa, **B**:llä on vain kaksi positiivista ominaisarvoa. Loput ovat kauniisti nollia laskutarkkuuden rajoissa. Matriisi **B** on siis tässä tapauksessa ei-negatiivisesti definiitti ja alkuperäinen etäisyysmatriisi euklidinen.

Jotta saisimme matriisia **B** vastaavan konfiguraatiomatriisin **X** (seuraavassa X2) joudumme kuten pääkomponenttianalyysissa kertomaan ominaisvektorit vastaavien ominaisarvojen neliöjuurilla:

```
 18  1 SURVO 84C EDITOR Wed Jun 08 09:17:53 1994        D:\M\MEN2\ 240 100 0
 51 *
 52 *MAT L=L(1:2,*)
 53 *MAT TRANSFORM L BY sqrt(X#)
 54 *MAT L=DV(L)        / *L~DV(T(L_by_sqrt(X#))) D2*2
 55 *MAT S=S(*,1:2)
 56 *MAT X2=S*L          / *X2~S(*,1:2)*DV(T(L_by_sqrt(X#))) 6*2
 57 *
 58 *MAT LOAD X2,CUR+1_
 59 *MATRIX X2
 60 *S(*,1:2)*DV(T(L_by_sqrt(X#)))
 61 *///           ev1      ev2
 62 *X1         5.81633  0.55610
 63 *X2         0.70529  3.36275
 64 *X3         0.85821 -0.63432
 65 *X4         0.97290 -3.63213
 66 *X5        -4.13813 -0.82547
 67 *X6        -4.21459  1.17307
 68 *
```

Tulos (**B**:n keskistyksestä johtuen) on valmiiksi keskistetty niin, että painopiste tulee origoon. Jos teemme vastaavan keskistyksen alkuperäiselle konfigu-

raatiomatriisille **X**, emme saa suoraan samaa esitystä:

```
 18  1 SURVO 84C EDITOR Wed Jun 08 09:21:49 1994        D:\M\MEN2\ 240 100 0
 68 *
 69 *MAT XC=CENTER(X)   / *XC~CENTER(X) 6*2
 70 *MAT LOAD XC,CUR+1_
 71 *MATRIX XC
 72 *CENTER(X)
 73 *///             A1       A2
 74 *  1       -5.83333  0.33333
 75 *  2       -0.83333  3.33333
 76 *  3       -0.83333 -0.66667
 77 *  4       -0.83333 -3.66667
 78 *  5        4.16667 -0.66667
 79 *  6        4.16667  1.33333
 80 *
```

Pistekonfiguraatio on kuitenkin täsmälleen alkuperäinen. Se on vain kiertynyt pääakseliesitykseen. Tämä vastaavuus saadaan helposti vahvistetuksi symmetrisen transformaatioanalyysin avulla:

```
 38  1 SURVO 84C EDITOR Wed Jun 08 09:27:42 1994        D:\M\MEN2\ 240 100 0
 80 *
 81 */TRAN-SYMMETR X2,XC
 82 *MAT LOAD L.M,##.###,END+2  / Transformation matrix
 83 *MAT LOAD E.M,##.################,END+2_ / Residual matrix
 84 *
 85 *MATRIX L.M
 86 *Transformation_matrix
 87 *///          A1     A2
 88 *ev1      -0.999 -0.038
 89 *ev2      -0.038  0.999
 90 *
 91 *MATRIX E.M
 92 *Residual_matrix
 93 *///                      A1                  A2
 94 *X1        0.000000000000000  0.000000000000000
 95 *X2       -0.000000000000001  0.000000000000001
 96 *X3       -0.000000000000000 -0.000000000000000
 97 *X4        0.000000000000001 -0.000000000000004
 98 *X5       -0.000000000000004 -0.000000000000001
 99 *X6        0.000000000000002  0.000000000000000
100 *
```

Sama skaalaustehtävä suoritetaan Survossa etäisyysmatriisista suoraan sukrolla `/CSCAL` seuraavasti. Etäisyysmatriisi **D** lasketaan tässä `DIST`-operaatiolla "havainnoista" `X.MAT` (yleensä lähtökohtana on aito havaintoaineisto).

```
 23  1 SURVO 84C EDITOR Sun Nov 06 12:38:03 1994          D:\M\MEN2\ 240 100 0
100 *.................................................................
101 *
102 *DIST X.MAT,D  / Etäisyysmatriisi D suoraan "havainnoista" X.MAT
103 *
104 */CSCAL D,2
105 *Classical multidimensional scaling for D:
106 *MAT LOAD CSCAL.M,END+2_/ Scale values (2 dimensions)
107 *MAT LOAD CSEIGEN.M,END+2 / Eigenvalues
108 *MAT LOAD CSCENT.M,END+2 / Eigenvalues (percentages)
109 *MAT LOAD CSDIST.M,END+2 / Reproduced distances
110 *GPLOT CSCAL.M,DIM1,DIM2 / POINT=[SMALL],CASE
111 *LSCAL D,CSCAL.M,END+2 / Least Squares Scaling
112 *
113 *MATRIX CSCAL.M
114 *CS_scales
115 *///          DIM1     DIM2
116 *X1        5.81633  0.55610
117 *X2        0.70529  3.36275
118 *X3        0.85821 -0.63432
119 *X4        0.97290 -3.63213
120 *X5       -4.13813 -0.82547
121 *X6       -4.21459  1.17307
122 *
```

Sukrokomennossa (rivi 104) viitataan vain etäisyysmatriisiin (D) ja haluttuun dimensioon (2). Sukro /CSCAL kirjoittaa tulostiedot komentorivin alapuolelle (tässä riveille 105-111). Ne sisältävät valmiita komentoja, joilla saadaan näkyviin halutut tulokset. Esim. rivin 106 komennolla tulostetaan konfiguraatiomatriisi, joka on täsmälleen sama kuin edellä matriisikäskyillä laskettu.

Muissa matriisitiedostoissa ovat ominaisarvot (CSEIGEN.M) ja näiden prosentuaaliset ja kumulatiiviset "selitysosuudet". Jos etäisyysmatriisi ei ole euklidinen, myös tästä tulee huomautus omalle rivilleen. Tällöin selitysosuudet lasketaan ominaisarvojen itseisarvoista. Matriisitiedostossa CSDIST.M ovat konfiguraatiomatriisista tai itse asiassa suoraan matriisista **B** kaavalla (1) lasketut etäisyydet, joita voi jälkikäteen verrata todellisiin etäisyyksiin esim. laskemalla matriisitulkilla erotuksen D - CSDIST.M . Tässä tapauksessa se on turhaa, koska yhteensopivuus oli täydellinen ja ko. erotus on nollamatriisi.

Mukana on lisäksi valmis GPLOT-komento, jolla piirretään analyysin tuottama "kartta". Analyysia voi jatkaa (rivin 111) LSCAL-komennolla, joka lähtien klassisen skaalauksen tuloksesta etsii iteroimalla pienimmän neliösumman ratkaisua. Tämäkin on tässä yhteydessä aiheetonta, koska paras mahdollinen tulos on jo saavutettu.

9.2 Pienimmän neliösumman skaalaus

Silloin kun etäisyysmatriisi ei ole puhtaasti euklidinen tai kun etäisyystiedoissa on puutteita tai suurempaa epävarmuutta, klassinen skaalaus ei välttämättä anna hyviä tuloksia. Klassisen skaalauksen tulos on kyllä optimaalinen esim. pienimmän neliösumman kriteerin mielessä, kun asiaa tarkastellaan edellä esitetyn etäisyysmatriisin muunnoksen **B** kautta, jolloin palaudutaan pääkomponenttianalyysiin. Tämä tulos ei kuitenkaan ole optimaalinen alkuperäisessä etäisyysmetriikassa, eli verrattaessa annettuja etäisyyksiä klassisen skaalauksen tuloksista laskettuihin saattaa esiintyä yllättävän suuria poikkeamia. Kuitenkin käytännössä eräs tärkeimpiä tarkistimia skaalauksen onnistumiselle on hyvä vastaavuus etäisyyksien kesken.

Pienimmän neliösumman skaalauksessa onkin tavoitteena yksinkertaisesti minimoida havaittujen ja estimoidusta konfiguraatiosta laskettujen etäisyyksien poikkeamien (painotettu) neliösumma eli minimoitava suure on muotoa

$$f(\mathbf{X}) = \sum_{r \neq s} w_{rs} [d_{rs} - \sqrt{(x_{r1} - x_{s1})^2 + (x_{r2} - x_{s2})^2 + ... + (x_{rp} - x_{sp})^2}\,]^2 ,$$

missä painot w_{rs} ovat tavallisesti joko ykkösiä tai w_{rs}=1/d_{rs} . Minimointi tapahtuu kaikkien **X**-matriisin alkioiden x_{ij} suhteen, joita on *np* kappaletta. Jos pyritään mahdollisimman yksikäsitteiseen ratkaisuun, näillä muuttujilla on *p*+*p*(*p*-1)/2 side-ehtoa, joista *p* liittyy keskistys- ja *p*(*p*-1)/2 rotaatiomahdollisuuteen.

Funktion *f*(**X**) rakenteesta ja em. side-ehdoista johtuen tehtävä ei käytännössä ole helppo, sillä varsinkin ristiriitaisia etäisyystietoja sisältävissä sovelluksissa funktiolla *f*(**X**) on lukuisia paikallisia minimikohtia ja satulapisteitä, jotka hankaloittavat globaalin minimikohdan löytämistä. Luonnollisesti hyvät alkuarvot auttavat ja tavallisesti lähdetään liikkeelle klassisen skaalauksen tuloksesta.

Survossa pienimmän neliösumman skaalaus tehdään `LSCAL`-operaatiolla, jonka rakenne on

`LSCAL` <etäisyysmatriisi>,<lähtökonfiguraatiomatriisi>,`L` .

Tämä operaatio sallii etäisyyksien painotuksen ohella (täsmennys `WEIGHT=` <painojen w_{rs} matriisi>) useita vaihtoehtoisia tapoja sekä etäisyysmetriikan (täsmennys `METRICS=Lp`, oletuksena `L2`) että yhteensopivuuskriteerin (täsmennys `CRITERION=Lp`, oletuksena `L2`) valintaan.

On myös mahdollista muuntaa etäisyyksiä d_{rs} additiivisella vakiolla, jonka lopulliseksi arvoksi tulee se, joka yhdessä **X**:n kanssa minimoi funktion *f*(**X**). Jos additiivista vakiota halutaan käyttää, sen alkuarvo annetaan täsmennyksellä `CONSTANT`.

Puuttuvat etäisyydet osoitetaan negatiivisilla arvoilla (esim. -1). Jos pistei-

den lukumäärä *n* on suuri, puuttuvia tietoja voi olla runsaastikin. Ainoa olennainen vaatimus on se, että tietoja on niin paljon, että konfiguraatiosta tulee jäykkä.

LSCAL-operaatio käyttää optimiratkaisua etsiessään *Nelder*in ja *Mead*in 1963 esittämää polytope-algoritmia, jossa minimoitavan funktion riippuessa *m* muuttujasta, *m*+1 pisteen muodostamaa simpleksiä kuljetetaan ja muunnetaan simpleksin kärkipisteissä saatujen arvojen mukaisesti. Koska tämä menetelmä ei käytä mitään funktion derivaattoihin perustuvaa tietoa, se ei aivan helposti lankea esim. satulapisteisiin ja se saattaa harpata yli paikallisten minikohtien. Heikkoutena on hitaus varsinkin suurilla muuttujamäärillä. Alkuperäistä simpleksiä muodostaessaan se käyttää STEP-täsmennyksellä ilmoitettua askelpituutta. Oletus on STEP=1. Koska usean askeleen jälkeen simpleksi saattaa surkastua, LSCAL palaa tarvittaessa alkuperäiseen askelpituuteen ja jatkaa parhaasta tähän asti löydetystä pisteestä.

Ohjelman rajoituksena on $np \leq 90$ eli 1-ulotteinen skaalaus voidaan tehdä 90 pisteellä, 2-ulotteinen 45 pisteellä ja 3-ulotteinen 30 pisteellä.

Tuloksena saadaan konfiguraatiomatriisi LSCAL.M, joka on pystyriveittäin keskistetty ja rotatoitu pääakselimuotoon. Verrattaessa eri aineistoilla saatuja tuloksia keskenään on silti syytä käyttää /TRAN-SYMMETR -sukroa.

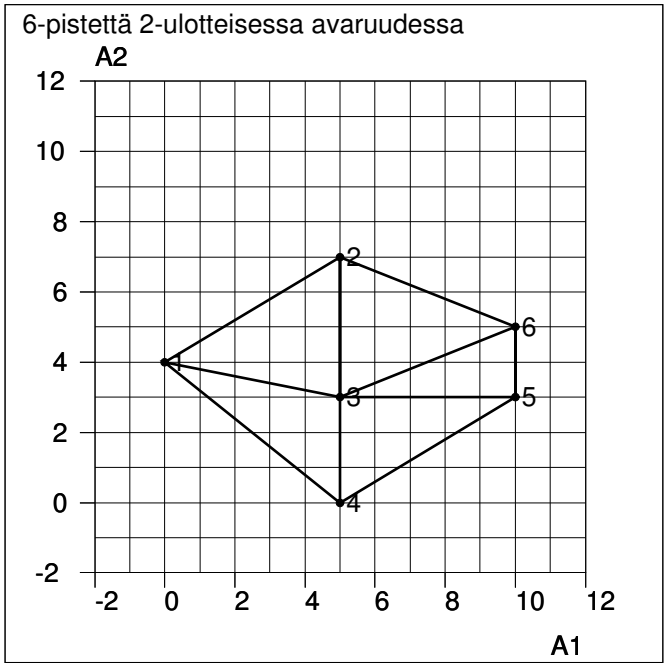
Ohjelma laskee lisäksi estimoitujen etäisyyksien matriisin LSDIST.M , jota voi suoraan verrata alkuperäiseen etäisyysmatriisiin ja tarkkailla erityisesti suuria poikkeamia. Puuttuvat etäisyydet näkyvät edelleen LSDIST.M-matriisissa samoina negatiivisina lukuina, joten esim. erotusmatriisia laskettaessa ne tulevat nolliksi.

9.2.1 Esimerkki 1 (jatkoa)

"Vaikeutamme" 6 pisteen esimerkkiämme asettamalla eräät etäisyydet tuntemattomiksi seuraavasti:

```
  1  1 SURVO 84C EDITOR Fri Jun 10 17:22:21 1994        D:\M\MEN2\ 300 100 0
122 *.....................................................................
123 *MAT D(1,5)=-1
124 *MAT D(5,1)=-1
125 *MAT D(1,6)=-1
126 *MAT D(6,1)=-1
127 *MAT D(2,4)=-1
128 *MAT D(4,2)=-1
129 *MAT D(2,5)=-1
130 *MAT D(5,2)=-1
131 *MAT D(4,6)=-1
132 *MAT D(6,4)=-1
133 *_
```

Tunnetuiksi jäävät seuraavaan kuvaan paksummalla viivalla merkityt yhteydet.



Tässä tapauksessa klassista skaalausta ei voi käyttää, koska se edellyttää kaikki etäisyydet annetuiksi. Jos esim. yrittää sijoittaa tuntemattomien etäisyyksien paikoille jonkinlaisia arvauksia ja sitten iteroimalla klassista skaalausta tarkentaa tulosta, se ei näytä onnistuvan.

Pienimmän neliösumman skaalaus sen sijaan toimii tietojen puuttumisesta huolimatta.

```
 23  1 SURVO 84C EDITOR Fri Jun 10 17:31:25 1994        D:\M\MEN2\ 300 100 0
133 *
134 *MATRIX X0
135 *///  X1 X2
136 *1    0  5
137 *2    6  6
138 *3    7  0
139 *4    4  1
140 *5    9  2
141 *6    8  6
142 *
143 *MAT SAVE X0
144 *
145 *LSCAL D,X0,CUR+1
146 *Least-squares scaling for 6*6 dissimilarity (distance) matrix D:
147 *Initial criterion value 77.9255  Dimension=2
148 *Final criterion value 2.03062e-018  nf=1401
149 *MAT LOAD LSCAL.M,END+2_ /  Solution in 2 dimensions
150 *MAT LOAD LSDIST.M,END+2  /  Estimated distances
151 *GPLOT LSCAL.M,X1,X2 / POINT=[SMALL],CASE
152 *
153 *MATRIX LSCAL.M
154 *LS_scaled
155 *///            X1       X2
156 *  1      -5.81633  0.55610
157 *  2      -0.70529  3.36275
158 *  3      -0.85821 -0.63432
159 *  4      -0.97290 -3.63213
160 *  5       4.13813 -0.82547
161 *  6       4.21459  1.17307
162 *
```

Lähtökonfiguraatioksi on valittu matriisi X0, joka vain etäisesti muistuttaa oikeata (matriisi X riveillä 2-9). Rivillä 145 on käynnistetty LSCAL-komento, joka viittaa nyt puuttuvia tietoja sisältävään etäisyysmatriisiin D ja lähtökonfiguraatioon X0.

LSCAL-operaation suoraan antamat tulokset ovat riveillä 146-151. Riviltä 147 näkyy, että tunnettujen etäisyyksien ja niitä vastaavien alkuasetelmasta laskettujen etäisyyksien erotusten neliösumma on 77.9255 . Tämä kokonaispoikkeama on huvennut 1401 neliösummakokeilun jälkeen käytännössä nollaan (2.03062e-018) eli on saatu rakenne, jossa etäisyydet sopivat täydellisesti yhteen annettujen etäisyyksien kanssa.

Kuten riveiltä 153-161 voi todeta, konfiguraatio on sama kuin klassisella skaalauksella täydellisistä etäisyystiedoista (riveillä 58-67) saatu. Täysin merkityksetöntä on ensimmäisen dimension etumerkkien kääntyminen.

Annettu rakenne ei kuitenkaan ole puuttuvista tiedoista johtuen aivan jäykkä, sillä piste 1, koordinaatteina (0,4) voidaan peilata pisteitä 2,3 ja 4 yhdistävän suoran toiselle puolelle pisteiden 5 ja 6 väliin pisteeksi (10,4) etäisyyksien lainkaan muuttumatta.

Niinpä, jos käytetään tämänsukuista lähtöasetelmaa XA, saadaan tämä vaihtoehtoinen kuvio:

```
 17  1 SURVO 84C EDITOR Fri Jun 10 17:59:11 1994          D:\M\MEN2\ 300 100 0
162 *.....................................................................
163 *MATRIX XA
164 *///  X1 X2
165 *1   11  2
166 *2    6  8
167 *3    4  4
168 *4    6  1
169 *5    9  2
170 *6   11  4
171 *
172 *MAT SAVE XA
173 *
174 *LSCAL D,XA,CUR+1_
175 *Least-squares scaling for 6*6 dissimilarity (distance) matrix D:
176 *Initial criterion value 45.1309  Dimension=2
177 *Final criterion value 1.07581e-019  nf=1423
178 *MAT LOAD LSCAL.M,END+2  /  Solution in 2 dimensions
179 *MAT LOAD LSDIST.M,END+2  /  Estimated distances
180 *GPLOT LSCAL.M,X1,X2 / POINT=[SMALL],CASE
181 *
```

Kuvassa konfiguraatio (`LSCAL.M`) näyttää seuraavalta

eli vastaa odotettua.

Kohdefunktiolla on myös huonompia paikallisia minimikohtia, kuten näkyy seuraavasta yrityksestä:

```
 17  1 SURVO 84C EDITOR Fri Jun 10 18:11:03 1994          D:\M\MEN2\ 300 100 0
181 *.....................................................................
182 *MATRIX XB
183 *///  X1 X2
184 *1    1  6
185 *2    2  5
186 *3    3  4
187 *4    4  3
188 *5    5  2
189 *6    6  1
190 *
191 *MAT SAVE XB
192 *
193 *LSCAL D,XB,CUR+1_
194 *Least-squares scaling for 6*6 dissimilarity (distance) matrix D:
195 *Initial criterion value 128.955  Dimension=2
196 *Final criterion value 0.338316  nf=1572
197 *MAT LOAD LSCAL.M,END+2  /  Solution in 2 dimensions
198 *MAT LOAD LSDIST.M,END+2  /  Estimated distances
199 *GPLOT LSCAL.M,X1,X2 / POINT=[SMALL],CASE
200 *
```

Lähtömatriisilla XB syntyy tulos, joka ei aivan täsmää annetun etäisyysmatriisin kanssa. Poikkeamat ovat tasaisesti nollan ympärillä eli kyseessä on suhteellisen hyvä approksimaatio:

```
 17  1 SURVO 84C EDITOR Fri Jun 10 20:07:45 1994          D:\M\MEN2\ 300 100 0
200 *
201 *MAT E!=D-LSDIST.M
202 *MAT LOAD E,CUR+1_
203 *MATRIX E
204 *///            X1       X2       X3       X4       X5       X6
205 *X1        0.00000 -0.13504 -0.04414  0.17122  0.00000  0.00000
206 *X2       -0.13504  0.00000  0.12281  0.00000  0.00000 -0.12644
207 *X3       -0.04414  0.12281  0.00000 -0.16736 -0.11947  0.11742
208 *X4        0.17122  0.00000 -0.16736  0.00000  0.15742  0.00000
209 *X5        0.00000  0.00000 -0.11947  0.15742  0.00000 -0.08797
210 *X6        0.00000 -0.12644  0.11742  0.00000 -0.08797  0.00000
211 *
```

Kun piirretään saatu konfiguraatio LSCAL.M,

havaitaan, että ratkaisu on muodostunut "nyrjäyttämällä" piste 4 "suoran" 1,3,5 "yli". Koska pisteet 1,3 ja 5 eivät ole aivan samalla suoralla, ratkaisu voi olla vain likimääräinen.

Tämä esimerkki kaikessa yksinkertaisuudessaan osoittaa, millaisia ongelmia moniulotteisessa skaalauksessa tulee vastaan, kun etäisyystiedot ovat puutteellisia tai ristiriitaisia.

9.2.2 Esimerkki 2

Toisena esimerkkinä tutkimme, kuinka hyvin onnistuu Suomen eräiden paikkakuntien sijoittelu kartalle tiedossa olevien maantie-etäisyyksien avulla. Teiden mutkikkuus saattaa tässä tapauksessa vaikeuttaa tehtävää niin, ettei pisteitä saakaan puhtaasti sjoitettua 2-ulotteiselle kartalle.

Autoilijan tiekartassa (1989) on eräiden paikkakuntien maantie-etäisyydet taulukkona, josta on muodostettu seuraava matriisi:

```
 16  1 SURVO 84C EDITOR Sat Jun 11 12:29:42 1994        D:\M\MEN2\ 200 100 0
  1 *
  2 *MATRIX SUOMI
  3 *///      Hels Jyvk Kari Kilp Kuus Nuor Oulu Rova Tamp Torn Turk Vaal Vaas
  4 *Helsinki    0  272 1262 1206  814 1334  612  837  174  744  164  184  419
  5 *Jyvkylä   272    0  988  933  553 1060  339  563  148  471  304  279  282
  6 *Karigasn 1262  988    0  402  546  144  650  425 1137  549 1283 1268  968
  7 *Kilpisj  1206  933  402    0  620  602  594  428 1081  461 1228 1212  912
  8 *Kuusamo   814  553  546  620    0  617  215  191  702  315  848  744  533
  9 *Nuorgam  1334 1060  144  602  617    0  721  497 1209  621 1355 1339 1040
 10 *Oulu      612  339  650  594  215  721    0  224  487  132  633  617  318
 11 *Rovanmi   837  563  425  428  191  497  224    0  712  124  858  842  543
 12 *Tampere   174  148 1137 1081  702 1209  487  712    0  620  155  276  241
 13 *Tornio    744  471  549  461  315  621  132  124  620    0  766  750  450
 14 *Turku     164  304 1283 1228  848 1355  633  858  155  766    0  345  331
 15 *Vaalimaa  184  279 1268 1212  744 1339  617  842  276  750  345    0  516
 16 *Vaasa     419  282  968  912  533 1040  318  543  241  450  331  516    0
 17 *
 18 *MAT SAVE SUOMI_
 19 *
```

Teemme klassisen skaalauksen /CSCAL-sukrolla:

```
 15  1 SURVO 84C EDITOR Sat Jun 11 12:33:42 1994        D:\M\MEN2\ 200 100 0
 19 *
 20 */CSCAL SUOMI,2_
 21 *Classical multidimensional scaling for SUOMI:
 22 *MAT LOAD CSCAL.M,END+2 / Scale values (2 dimensions)
 23 *MAT LOAD CSEIGEN.M,END+2 / Eigenvalues
 24 *MAT LOAD CSCENT.M,END+2 / Eigenvalues (percentages)
 25 *MAT LOAD CSDIST.M,END+2 / Reproduced distances
 26 *GPLOT CSCAL.M,DIM1,DIM2 / POINT=[SMALL],CASE
 27 *LSCAL SUOMI,CSCAL.M,END+2 / Least Squares Scaling
 28 *Distance matrix SUOMI is not Euclidean!
 29 *
```

Riviltä 28 huomaamme, että etäisyysmatriisi ei ole euklidinen. Tämä näkyy **B**-matriisin ominaisarvoista:

```
 22  1 SURVO 84C EDITOR Sat Jun 11 12:43:00 1994        D:\M\MEN2\ 200 100 0
 29 *
 30 *MAT CSCENT=CSCENT.M'  / Transponoidaan, jotta matriisi näkyisi kokonaan
 31 *MAT LOAD CSCENT,CUR+1_/ toimituskentässä.
 32 *MATRIX CSCENT
 33 *Eigenvalues_(in_percentages)'
 34 *///      Per_cent Cumulat.
 35 *DIM1       84.649   84.649
 36 *DIM2        5.889   90.539
 37 *DIM3        3.854   94.393
 38 *DIM4        1.803   96.195
 39 *DIM5        0.382   96.577
 40 *DIM6        0.143   96.720
 41 *DIM7        0.010   96.730
 42 *DIM8       -0.000   96.730
 43 *DIM9       -0.013   96.743
 44 *DIM10      -0.181   96.924
 45 *DIM11      -0.524   97.448
 46 *DIM12      -0.973   98.420
 47 *DIM13      -1.580  100.000
 48 *
```

Ensimmäinen dimensio on todella vahva. Huomaamme kohta, että se vastaa likimain pohjois-etelä-suuntaa. Toinen dimensio on huomattavasti heikompi ja sitä lähentelee kolmas. Tämä on seurausta siitä, että Suomessa idän ja lännen väliset tieyhteydet esim. vesistöistä johtuen ovat mutkaisemmat kuin pohjois-etelä-suunnassa. Toinen luonnollinen päädimensio ikäänkuin hajoaa useammalle komponentille. Kuitenkin kahden ensimmäisen ulottuvuuden yhteinen selitysosuus (90.5 %) on varsin tyydyttävä.

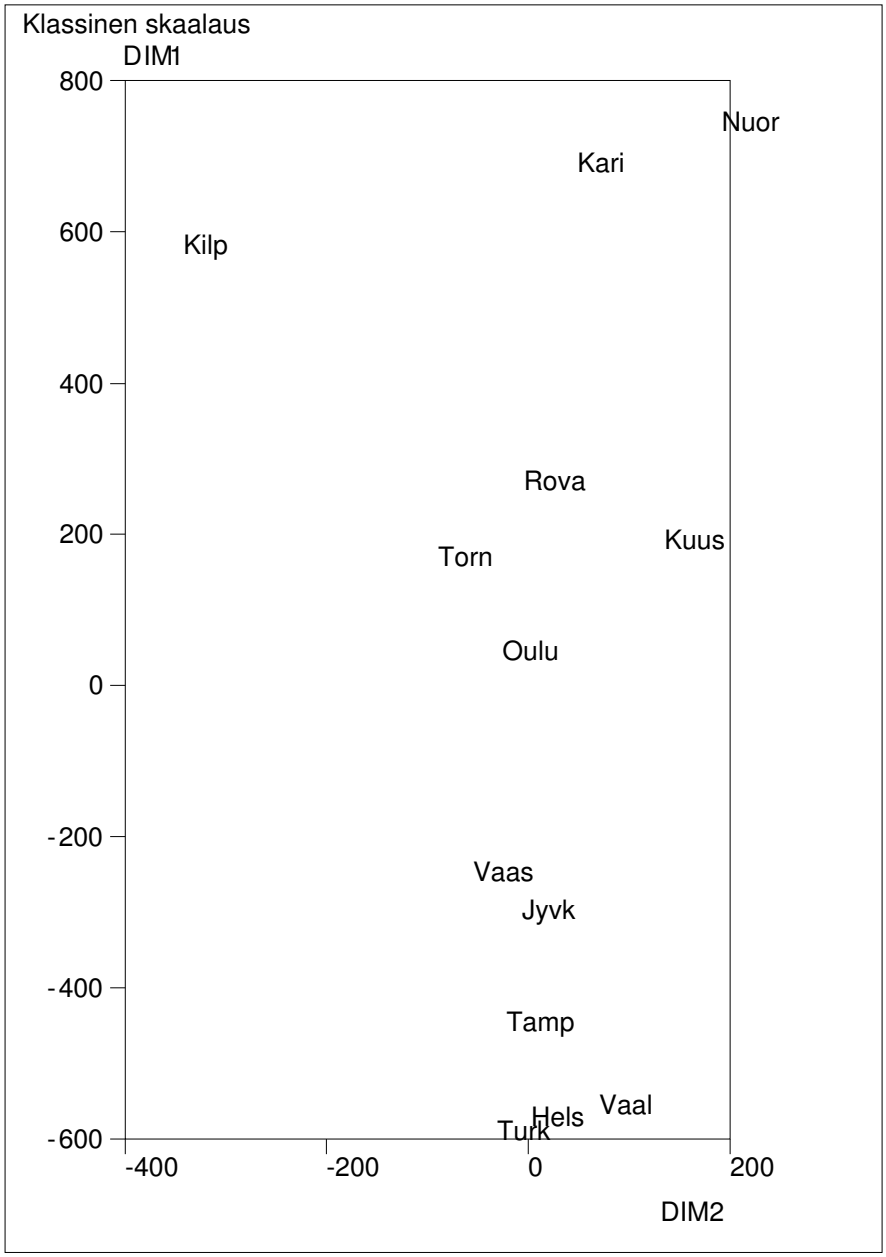
Viisi viimeistä ominaisarvoa ovat negatiivisia. Yksi **B**:n ominaisarvo on (keskistyksistä johtuen) aina tarkalleen 0. Tässä tapauksessa se on kahdeksas.

Jos katsomme, miltä näyttävät etäisyyksien poikkeamat alkuperäisen etäisyysmatriisin `SUOMI` ja konfiguraatiosta laskettujen (`CSDIST.M`) välillä,

```
 21  1 SURVO 84C EDITOR Sat Jun 11 12:57:33 1994        D:\M\MEN2\ 200 100 0
 48 *
 49 *MAT E!=SUOMI-CSDIST.M        / *E~SUOMI-CS_distances S13*13
 50 *MAT LOAD E,###,CUR+1_
 51 *MATRIX E
 52 *///      Hel Jyv Kar Kil Kuu Nuo Oul Rov Tam Tor Tur Vaa Vaa
 53 *Helsinki   0  -1  -0   3  39   3  -6  -5  46  -2 126 115  90
 54 *Jyvkylä   -1   0  -2  -8  43  -4  -5  -5  -1  -4  11  10 212
 55 *Karigasn  -0  -2   0  -4  40 -24   1   2  -1  10   1  21  25
 56 *Kilpisj    3  -8  -4   0   5  29 -28 -31   5 -22  17   2  35
 57 *Kuusamo   39  43  40   5   0  60  -2  31  45  90  49  -7  55
 58 *Nuorgam    3  -4 -24  29  60   0 -15 -22  -3 -26  -0  31  14
 59 *Oulu      -6  -5   1 -28  -2 -15   0  -2  -5  -7  -2   8  23
 60 *Rovanmi   -5  -5   2 -31  31 -22  -2   0  -4  -8  -2  12  23
 61 *Tampere   46  -1  -1   5  45  -3  -5  -4   0   1  11 132  40
 62 *Tornio    -2  -4  10 -22  90 -26  -7  -8   1   0   5   7  32
 63 *Turku    126  11   1  17  49  -0  -2  -2  11   5   0 238 -12
 64 *Vaalimaa 115  10  21   2  -7  31   8  12 132   7 238   0 183
 65 *Vaasa     90 212  25  35  55  14  23  23  40  32 -12 183   0
 66 *
```

on todettava, ettei klassinen skaalaus ole toiminut kunnolla etenkään Etelä-Suomen kohdalla. Esim. Helsingin ja Turun välinen etäisyydessä, 164 km, on virhettä 126 km eli ko. etäisyys olisi näiden tulosten mukaan vain 38 km!

Parhaiten skaalauksen epäonnistuminen näkyy kartalta, joka on piirretty rivin 26 `GPLOT`-komentoa vastaavalla Survon piirroskaaviolla. Luonnollista karttaesitystä tavoiteltaessa on syytä vaihtaa dimensioitten järjestys:



Koko eteläisin Suomi on kutistunut ja pohjoinen leventynyt. Eräs syy lienee mm. Nuorgamin ja Kilpisjärven välinen kohtuuttoman hankala tieyhteys.

On aihetta kokeilla, pystyykö pienimmän neliösumman skaalaus parantamaan tätä tulosta. Käynnistämme siis `LSCAL`-operaation niin, että lähtökonfiguraatio on klassisen skaalauksen tuottama `CSCAL.M`.

```
 26  1 SURVO 84C EDITOR Sat Jun 11 15:14:15 1994        D:\M\MEN2\ 200 100 0
 67 *.....................................................................
 68 *LSCAL SUOMI,CSCAL.M,CUR+1_
 69 *Least-squares scaling for 13*13 dissimilarity (distance) matrix SUOMI:
 70 *Initial criterion value 461546  Dimension=2
 71 *Final criterion value 78536.8  nf=12334
 72 *MAT LOAD LSCAL.M,END+2  /  Solution in 2 dimensions
 73 *MAT LOAD LSDIST.M,END+2  /  Estimated distances
 74 *GPLOT LSCAL.M,DIM1,DIM2 / POINT=[SMALL],CASE
 75 *
```

Rivillä 68 aktivoidun LSCAL-komennon suoraan antamat tulokset ovat riveillä 69-74. Riveiltä 70 ja 71 voi todeta, että virheneliösumma on pienentynyt klassisen skaalauksen mukaisesta alkutilanteesta niin paljon, että lopullinen neliösumma on vain noin 17% siitä. Ratkaisu on ollut melko raskaan työn takana, sillä minimoitava kohdefunktio (neliösumma) on jouduttu laskemaan peräti 12334 kertaa, mikä nykyisillä laitteistoilla sujuu kyllä nopeasti.

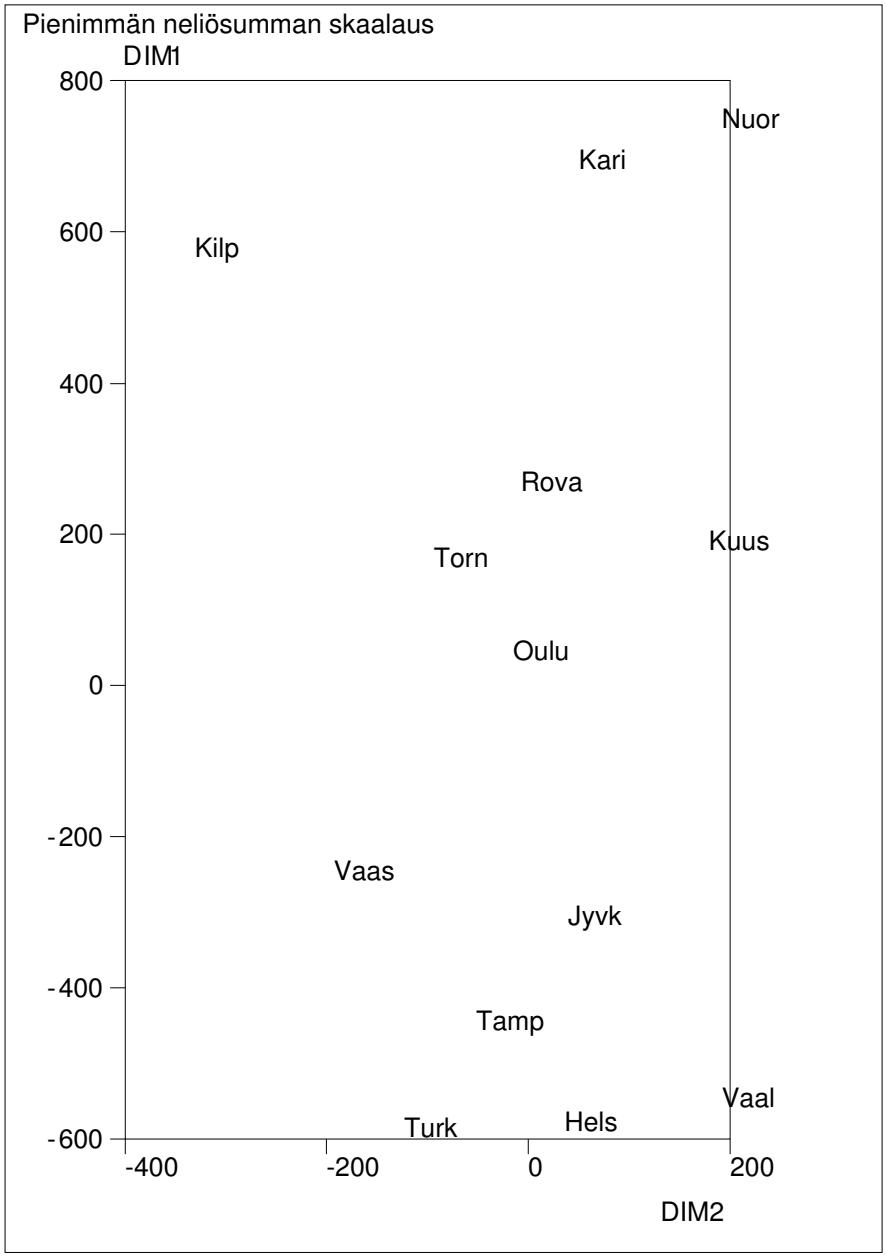
Laskennan aikana LSCAL tulostaa väliaikaiseen ikkunaan tietyin välein kohdefunktion laskentakertojen lukumäärän (nf), pienimmän siihen asti saavutetun neliösumman (f) ja suhteellisen muutoksen edellisestä neliösummasta. Käyttäjällä on tilaisuus keskeyttää toiminta painamalla pistettä tai jos suppeneminen kohti ratkaisua tuntuu hidastuvan simpleksin luhistuessa, aloittaa tuoreella virityksellä parhaasta siihenastisesta kohdasta napilla N. Muuten iterointikertoja säätelevät täsmennykset MAXNF ja EPS. Edellisen oletusarvo on 10000 ja se ilmoittaa, kuinka monta kertaa kohdefunktio lasketaan ilman simpleksin päivitystä. Simpleksi päivitetään myös silloin, kun suhteellinen muutos on alle EPS-täsmennyksellä ilmoitetun rajan, jonka oletusarvo on 1e-5 (siis 10^{-5}). Jos simpleksin päivityksen jälkeen tulos ei parane tietyn laskentamäärän aikana, laskenta päättyy automaattisesti. Tällöin on ilmeistä, että ainakin paikallinen ääriarvo on saavutettu.

Tässä esimerkissä tulos on saavutettu oletusarvoilla ja ilman manuaalisia välivaikutuksia. Yli 12000 kohdefunktion laskentakerrasta lähes 9000 on mennyt alkuperäisen simpleksin mukaiseen iterointiin. Vasta tällöin on EPS-raja alitettu. Simpleksi olisi voitu uudistaa (N-napilla) jo selvästi aikaisemmin, jolloin selvitään jopa puolella nyt kuluneesta laskentatyöstä. Myös jos valitaan MAXNF=3000, saadaan sama tulos jo 7003 kohdefunktion laskentakerralla.

Kun nyt tarkastellaan havaittujen ja estimoitujen etäisyyksien erotuksia,

```
 21  1 SURVO 84C EDITOR Sat Jun 11 15:57:02 1994          D:\M\MEN2\ 200 100 0
 75 *
 76 *MAT E!=SUOMI-LSDIST.M         / *E~SUOMI-LS_distances S13*13
 77 *MAT LOAD E,###,CUR+1_
 78 *MATRIX E
 79 *///       Hel Jyv Kar Kil Kuu Nuo Oul Rov Tam Tor Tur Vaa Vaa
 80 *Helsinki   0  -1 -11  -8  32  -2 -13 -10  14 -14   5  10  15
 81 *Jyvkylä   -1   0 -12 -26  37  -6 -15 -13 -18 -22 -20 -13  42
 82 *Karigasn -11 -12   0   4  26  -9  -3  -5  -7   4  -9  18  -4
 83 *Kilpisj   -8 -26   4   0 -21  51 -27 -20  20 -12  45 -34  76
 84 *Kuusamo   32  37  26 -21   0  59 -28 -11  26  41  15   7 -41
 85 *Nuorgam   -2  -6  -9  51  59   0 -14 -24  -9 -26 -16  45 -27
 86 *Oulu     -13 -15  -3 -27 -28 -14   0   0  -4 -15  -6 -13 -23
 87 *Rovanmi  -10 -13  -5 -20 -11 -24   0   0  -2  -8  -4   1  -4
 88 *Tampere   14 -18  -7  20  26  -9  -4  -2   0   5  -3  -1  -3
 89 *Tornio   -14 -22   4 -12  41 -26 -15  -8   5   0  11 -25  23
 90 *Turku      5 -20  -9  45  15 -16  -6  -4  -3  11   0  12 -16
 91 *Vaalimaa  10 -13  18 -34   7  45 -13   1  -1 -25  12   0  16
 92 *Vaasa     15  42  -4  76 -41 -27 -23  -4  -3  23 -16  16   0
 93 *
```

havaitaan tuloksen järkevöityneen kauttaaltaan. Esim. Helsingin ja Turun välinen etäisyys heittää vain 5 km. Suurin virhe on Vaasan ja Kilpisjärven keskinäisessä etäisyydessä (76 km). Tulos on kuvassakin paljon parempi:


Pienimmän neliösumman skaalaus
DIM1
800
600
400
200
0
-200
-400
-600
Nuor
Kari
Kilp
Rova
Kuus
Torn
Oulu
Vaas
Jyvk
Tamp
Vaal
Turk
Hels
-400
-200
0
200
DIM2

9.2.3 Esimerkki 3

Aikaisemmisssa esimerkeissä kohteet ovat edustaneet fysikaalista todellisuutta, jolloin ei ole epäilystäkään siitä, etteikö moniulotteisella skaalauksella saataisi järkevänsuuntaisia tuloksia.

Todellisissa sovelluksissa tavoitteena on löytää mielekkäitä rakenteita jopa täysin abstrakteista ilmiöistä. Lähtökohtana on esim. yhden tai useamman henkilön subjektiiviset arviot havaintokohteiden keskinäisistä samanlaisuuksista, erilaisuuksista tai etäisyyksistä.

Pyysin Olli Mustosta valitsemaan 10 huomattavaa säveltäjää musiikin eri aikakausilta ja vertaamaan näitä täysin intuitiivisesti heidän koko tuotantonsa ja tyylinsä pohjalta. Sovimme, että hän käyttää asteikkoa 0 - 100 siten, että mitä enemmän hän katsoo säveltäjien eroavan toisistaan, sitä suuremman pistemäärän hän antaa. Noin puolen tunnin harkinnan jälkeen hän esitti seuraavan etäisyysmatriisin:

```
 29  1 SURVO 84C EDITOR Sun Jun 12 09:07:20 1994        D:\M\MEN2\ 200 100 0
  1 *
  2 *MATRIX MUS
  3 *///      Bach Hayd Moza Beet Schu Brah Sibe Debu Bart Sost
  4 *Bach        0   50   30   20   40   40   40   50   30   30
  5 *Haydn      50    0   10   15   30   70   90   50   80   40
  6 *Mozart     30   10    0   20   25   40   70   50   80   50
  7 *Beethven   20   15   20    0   10   20   25   80   60   40
  8 *Schubert   40   30   25   10    0   15   60   50   70   60
  9 *Brahms     40   70   40   20   15    0   20   70   70   70
 10 *Sibelius   40   90   70   25   60   20    0   35   35   20
 11 *Debussy    50   50   50   80   50   70   35    0   15   40
 12 *Bartok     30   80   80   60   70   70   35   15    0   20
 13 *Sostakov   30   40   50   40   60   70   20   40   20    0
 14 *
 15 *MAT SAVE MUS
 16 *MAT MUST=MUS'                / Pitempien riviotsikoiden
 17 *MAT CLABELS FROM MUST TO MUS_ / kopiointi sarakeotsikoiksi
 18 *
```

Säveltäjät esiintyvät taulukossa suurin piirtein aikajärjestyksessä. O.M. käytti asteikkoa 5 yksikön välein, koska hän katsoi, ettei ole edellytyksiä tarkempaan arviointiin. Suurin etäisyys 90 esiintyy Sibeliuksen ja Haydnin välillä. Matriisista `MUS` tehtiin 2-ulotteinen klassinen skaalaus:

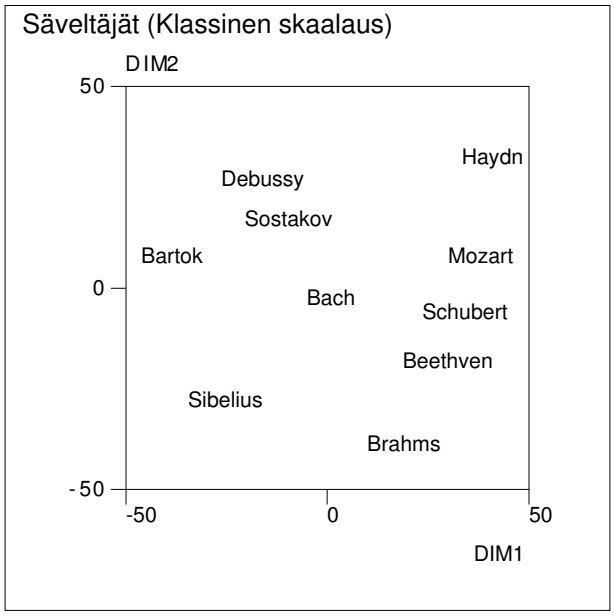
```
 22  1 SURVO 84C EDITOR Sun Jun 12 09:23:02 1994        D:\M\MEN2\ 200 100 0
 18 *
 19 */CSCAL MUS,2
 20 *Classical multidimensional scaling for MUS:
 21 *MAT LOAD CSCAL.M,END+2 / Scale values (2 dimensions)
 22 *MAT LOAD CSEIGEN.M,END+2 / Eigenvalues
 23 *MAT LOAD CSCENT.M,END+2 / Eigenvalues (percentages)
 24 *MAT LOAD CSDIST.M,END+2 / Reproduced distances
 25 *GPLOT CSCAL.M,DIM1,DIM2 / POINT=[SMALL],CASE
 26 *LSCAL MUS,CSCAL.M,END+2 / Least Squares Scaling
 27 *Distance matrix MUS is not Euclidean!
 28 *
 29 *MAT CSCENT=CSCENT.M'         / *CSCENT~Eigenvalues_(in_percentages)'
 30 *MAT LOAD CSCENT,CUR+1_
 31 *MATRIX CSCENT
 32 *Eigenvalues_(in_percentages)'
 33 *///      Per_cent Cumulat.
 34 *DIM1       35.090   35.090
 35 *DIM2       22.720   57.809
 36 *DIM3       10.765   68.574
 37 *DIM4        3.536   72.110
 38 *DIM5        2.416   74.526
 39 *DIM6       -0.000   74.526
 40 *DIM7       -3.112   77.637
 41 *DIM8       -4.260   81.897
 42 *DIM9       -4.405   86.302
 43 *DIM10     -13.698  100.000
 44 *
```

Kuten ominaisarvojen käyttäytymisestä näkyy, etäisyysmatriisi sellaisenaan ei ole euklidinen. Kaksi ensimmäistä ulottuvuutta kuitenkin selittävät lähes 60% koko vaihtelusta ja 77% "positiivisesta" vaihtelusta. Kohtalaiset negatiiviset ominaisarvot viittavat siihen, ettei arviointi voinut olla aivan ristiriidatonta. Sopivalla etäisyyksien epälineaarisella muunnoksella (esim. ottamalla neliöjuuri) matriisi tulisi lähemmäksi euklidista, mutta tällaisiin toimiin ei tässä tapauksessa ryhdytty.

Todellisten ja skaalauksen perusteella saatujen etäisyyksien poikkeamat ovat:

```
 21  1 SURVO 84C EDITOR Sun Jun 12 13:41:43 1994        D:\M\MEN2\ 240 100 0
 44 *
 45 *MAT E!=MUS-CSDIST.M          / *E~MUS-CS_distances S10*10
 46 *MAT LOAD E,###,CUR+1_
 47 *MATRIX E
 48 *///      Bac Hay Moz Bee Sch Bra Sib Deb Bar Sos
 49 *Bach       0  -2  -7  -8  11   1   1  14 -12   5
 50 *Haydn     -2   0 -15 -38 -10  -5  -1 -10  -3 -16
 51 *Mozart    -7 -15   0  -8  10 -11  -4  -9   4  -1
 52 *Beethven  -8 -38  -8   0  -3  -2 -29  16 -10 -13
 53 *Schubert  11 -10  10  -3   0 -20  -2 -10  -1  10
 54 *Brahms     1  -5 -11  -2 -20   0 -26  -5  -3   6
 55 *Sibelius   1  -1  -4 -29  -2 -26   0 -20  -3 -27
 56 *Debussy   14 -10  -9  16 -10  -5 -20   0 -12  29
 57 *Bartok   -12  -3   4 -10  -1  -3  -3 -12   0  -7
 58 *Sostakov   5 -16  -1 -13  10   6 -27  29  -7   0
 59 *
```

Huolimatta joistain suurista eroista, kuvassa tulos näyttää mielenkiintoiselta:

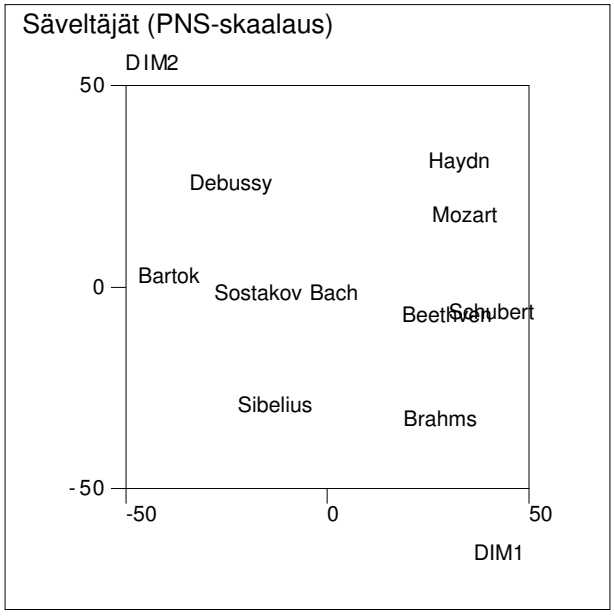


Ensimmäinen ulottuvuus oikealta vasemmalle vastaa muuten varsin hyvin aikaa vain sillä huomattavalla poikkeamalla, että "ajaton" Bach asettuu keskelle. Toinen dimensio on tulkittavissa ylhäältä alaspäin siirtymisenä "kevyestä raskaaseen" musiikkiin. Niinpä Wieniläisklassikot (Haydn, Mozart, Schubert ja Beethoven) muodostavat johdonmukaisen ketjun ja saavat jatkokseen vielä Brahmsin, joka Sibeliuksen kera sijoittuu "raskaimpaan sarjaan". Moderneimmat säveltäjät (Debussy, Shostakovits ja Bartok) muodostavat oman ryhmänsä ja on täysin ymmärrettävää, että näistä Shostakovits on lähinnä Bachia. Sibelius asettuu omaan yksinäisyyteensä.

Tämä asettelu ja tulkinta ei muutu, vaikka sovelletaan muita skaalaustapoja. Esim. pienimmän neliösumman skaalaus, kun käytetään additiivista vakiota alkuarvolla `CONSTANT=0`, tuottaa klassisesta skaalauksesta lähtien tulokset:

```
 21  1 SURVO 84C EDITOR Sun Jun 12 14:24:15 1994        D:\M\MEN2\ 240 100 0
 61 *.....................................................................
 62 *LSCAL MUS,CSCAL.M,CUR+1   /  CONSTANT=0  MAXNF=3000
 63 *Least-squares scaling for 10*10 dissimilarity (distance) matrix MUS:
 64 *Initial criterion value 16106.7  Dimension=2
 65 *Final criterion value 9551.86  nf=3388
 66 *Distance transformation D+3.96678
 67 *MAT LOAD LSCAL.M,END+2  /  Solution in 2 dimensions
 68 *MAT LOAD LSDIST.M,END+2  /  Estimated distances
 69 *GPLOT LSCAL.M,DIM1,DIM2 / POINT=[SMALL],CASE
 70 *
 71 *MAT E!=MUS-LSDIST.M          / *E~MUS-LS_distances S10*10
 72 *MAT C=CON(10,10,3.96678)
 73 *MAT D=IDN(10,10,3.96678)
 74 *MAT E=E+C           / *E~E+CON S10*10
 75 *MAT E!=E-D          / *E~E+CON-IDN S10*10
 76 *MAT LOAD E,###,CUR+1_
 77 *MATRIX E
 78 *///      Bac Hay Moz Bee Sch Bra Sib Deb Bar Sos
 79 *Bach       0  10  -2   0   9   5  11  13  -9  10
 80 *Haydn      10   0   1 -20  -4  10  17  -6   7 -18
 81 *Mozart     -2   1   0  -2   4  -7   6  -7  10  -3
 82 *Beethven    0 -20  -2   0   2  -2 -18  22  -2  -3
 83 *Schubert    9  -4   4   2   0 -10   7 -18  -4   6
 84 *Brahms      5  10  -7  -2 -10   0 -17  -5  -1  18
 85 *Sibelius   11  17   6 -18   7 -17   0 -18  -2  -4
 86 *Debussy    13  -6  -7  22 -18  -5 -18   0  -7  16
 87 *Bartok     -9   7  10  -2  -4  -1  -2  -7   0   5
 88 *Sostakov   10 -18  -3  -3   6  18  -4  16   5   0
 89 *
```

Additiivisen vakion lopullinen arvo on hieman alle 4 ja poikkeamataulukko on siistimpi kuin klassisessa skaalauksessa. Kuriositeettina mainittakoon, että suurin poikkeama (22) on Debussyn ja Beethovenin välillä. Tähän O.M. totesi vaikuttaneen sen, että Debussyn tiedetään vihanneen Beethovenin musiikkia. Näin hän sijoitti nämä säveltäjät kauemmaksi toisistaan, kuin mitä ilman tätä tietoa olisi tapahtunut.



10. Korrespondenssianalyysi

10.1. Määritelmä

Korrespondenssianalyysi (Correspondence Analysis) on sukua pääkomponenttianalyysille, mutta siinä käsitellään yleensä frekvenssitaulukoita havaintomatriisien asemasta. Korrespondenssianalyysia ovat harrastaneet erityisesti ranskalaiset. Se voidaan johtaa useista erilaisista lähtökohdista. Yksinkertaisin ja samalla luultavasti ensimmäinen perustelu liittyy kaksiulotteiseen skaalausongelmaan, jonka ratkaisuineen esitti *R.A.Fisher* vuonna 1940.

Fisher tarkasteli aineistoa, jossa on taulukoituna koululaisten (Caithness, Skotlanti) tukan ja silmien väri seuraavasti:

			SILMÄT		
	blue	light	medium	dark	*X*
TUKKA					
fair	326	688	343	98	-1.219
red	38	116	84	48	-0.523
medium	241	584	909	403	-0.094
dark	110	188	412	681	1.319
black	3	4	26	85	2.452
Y	-0.897	-0.987	0.075	1.574	

Sekä tukan että silmien värin luokitus on ainoastaan nominaaliasteikollista. Näyttää kuitenkin siltä, että tukan tummuusaste korreloi ainakin jonkin verran silmien tummuuden kanssa. Tarkoituksena on kvantifioida ko. korrelaatio antamalla sekä tukan että silmien värille mittaluvut X_1, X_2, X_3, X_4, X_5 ja Y_1, Y_2, Y_3, Y_4 siten, että muuttujien ylläolevasta luokitetusta aineistosta laskettu korrelaatiokerroin tulee mahdollisimman suureksi.

Fisherin tällä periaatteella laskemat mittaluvut on lisätty taulukkoon (sarake *X* ja rivi *Y*) ja maksimaalinen korrelaatiokerroin on 0.4464 .

Johdetaan nämä tulokset yleisesti tarkastelemalla $m \times n$-frekvenssitaulukkoa **F**, missä $m \geq n$:

	Y_1	Y_2	...	Y_n	Σ
X_1	f_{11}	f_{12}	...	f_{1n}	$f_{1.}$
X_2	f_{21}	f_{22}	...	f_{2n}	$f_{2.}$
...	...	...	...	...	...
X_m	f_{m1}	f_{m2}	...	f_{mn}	$f_{m.}$
Σ	$f_{.1}$	$f_{.2}$	...	$f_{.n}$	N

Koska asteikkoja X ja Y voi korrelaatiokertoimen säilyessä muuntaa lineaarisesti, oletetaan keskiarvot nolliksi ja varianssit ykkösiksi eli

$$(1)\quad \overline{X} = \frac{1}{N}\sum_{i=1}^{m} f_{i.} X_i = 0\ , \quad \overline{Y} = \frac{1}{N}\sum_{j=1}^{n} f_{.j} Y_j = 0$$

$$(2)\quad s_X^2 = \frac{1}{N}\sum_{i=1}^{m} f_{i.} X_i^2 = 1\ , \quad s_Y^2 = \frac{1}{N}\sum_{j=1}^{n} f_{.j} Y_j^2 = 1\ .$$

Tällöin korrelaatiokerroin r on yksinkertaisesti

$$(3)\quad r = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} X_i Y_j\ .$$

On siis maksimoitava lauseke (3) ehdoilla (1) ja (2).

Merkitään

$$(4)\quad \begin{aligned} U_i &= \sqrt{f_{i.}/N}\ X_i\ ,\ i = 1,2,...,m\ , \\ V_j &= \sqrt{f_{.j}/N}\ Y_i\ ,\ j = 1,2,...,n\ , \end{aligned}$$

jolloin yhtälöistä (2) seuraa, että $\|\mathbf{u}\|=1$ ja $\|\mathbf{v}\|=1$. Tällöin korrelaatiokerroin voidaan kirjoittaa vektorien $\mathbf{u}$ ja $\mathbf{v}$ avulla muotoon

$$r = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} (f_{ij} / \sqrt{f_{i.} f_{.j}})\ U_i V_j \ = \ \mathbf{u}'\mathbf{D}_m^{-1/2}\mathbf{F}\mathbf{D}_n^{1/2}\mathbf{v}\ ,$$

missä $\mathbf{D}_m$ on reunafrekvenssien $f_{1.}, f_{2.}, ..., f_{m.}$
ja $\mathbf{D}_n$ on reunafrekvenssien $f_{.1}, f_{.2}, ..., f_{.n}$
muodostama lävistäjämatriisi.

On siis maksimoitava $\mathbf{u}'\mathbf{A}\mathbf{v}$, missä $\mathbf{A}$ on $m\times n$-matriisi

$$\mathbf{A} = \mathbf{D}_m^{-1/2}\mathbf{F}\mathbf{D}_n^{1/2}$$

ehdoilla $\|\mathbf{u}\|=1$ ja $\|\mathbf{v}\|=1$. Optimiratkaisut löytyvät silloin matriisin $\mathbf{A}$ singulaariarvohajotelmasta

$$(5)\quad \mathbf{D}_m^{-1/2}\mathbf{F}\mathbf{D}_n^{1/2} = \mathbf{U}\mathbf{D}\mathbf{V}'$$

eli singulaariarvot $d_1, d_2, ..., d_n$ vastaavat maksimaalisia korrelaatiokertoimia.

Yhtälöiden (4) perusteella optimaaliset asteikot ovat

$$\begin{aligned} \mathbf{x}^{(i)} &= \sqrt{N}\,\mathbf{D}_m^{-1/2}\mathbf{u}^{(i)}\ , \\ \mathbf{y}^{(i)} &= \sqrt{N}\,\mathbf{D}_n^{1/2}\mathbf{v}^{(i)}\ ,\ i = 1,2,...,n\ . \end{aligned}$$

Suurin singulaariarvo on d_1=1 ja sitä vastaavat singulaarivektorit

$$\mathbf{u}^{(1)} = \mathbf{D}_m^{1/2}\mathbf{1}_m ,$$

$$\mathbf{v}^{(1)} = \mathbf{D}_n^{1/2}\mathbf{1}_n ,$$

missä esim. $\mathbf{1}_m$ on m ykkösen muodostama pystyvektori. Tämä todetaan osoittamalla, että $\mathbf{A}\mathbf{v}^{(1)}=d_1\mathbf{u}^{(1)}$ ja $\mathbf{A}'\mathbf{u}^{(1)}=d_1\mathbf{v}^{(1)}$. Esimerkiksi

$$\mathbf{A}\mathbf{v}^{(1)} = \mathbf{D}_m^{-1/2}\mathbf{F}\mathbf{1}_n = \mathbf{D}_m^{-1/2}\mathbf{D}_m\mathbf{1}_m = \mathbf{D}_m^{1/2}\mathbf{1}_m = \mathbf{u}^{(1)} = d_1\mathbf{u}^{(1)} .$$

Tämä ratkaisu antaisi esim.

$$\mathbf{x}^{(1)} = \sqrt{N}\,\mathbf{1}_m ,$$

mikä ei täytä ehtoa (1).

Muut ratkaisut

d_i , $\mathbf{u}^{(i)}$, $\mathbf{v}^{(i)}$, $i = 2,...,n$

täyttävät ehdon (1), sillä

$$0 = \sqrt{N}\,\mathbf{u}^{(1)'}\mathbf{u}^{(i)} = \sqrt{N}\,\mathbf{1}_m'\mathbf{D}_m^{1/2}(1/\sqrt{N})\mathbf{D}_m^{1/2}\mathbf{x}^{(i)} = \mathbf{1}_m'\mathbf{D}_m\mathbf{x}^{(i)} = \sum_{k=1}^{m} f_{k.}X_k^{(i)} .$$

Merkitsemällä

$$\mathbf{X} = [\,\mathbf{x}^{(1)}\,\mathbf{x}^{(2)}\,...\,\mathbf{x}^{(n)}\,] ,$$

$$\mathbf{Y} = [\,\mathbf{y}^{(1)}\,\mathbf{y}^{(2)}\,...\,\mathbf{y}^{(n)}\,] ,$$

voidaan kirjoittaa

$$\mathbf{X} = \sqrt{N}\,\mathbf{D}_m^{-1/2}\mathbf{U} ,$$

(6)

$$\mathbf{Y} = \sqrt{N}\,\mathbf{D}_n^{1/2}\mathbf{V} .$$

Koska hajotelma (5) voidaan esittää muodossa

(7) $\mathbf{F} = \mathbf{D}_m^{1/2}\mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{D}_n^{1/2}$,

saadaan

$$\mathbf{D}_m^{-1}\mathbf{F}\mathbf{Y} = \mathbf{D}_m^{-1}\mathbf{D}_m^{1/2}\mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{D}_n^{1/2}\sqrt{N}\,\mathbf{D}_n^{1/2}\mathbf{V} = \sqrt{N}\,\mathbf{D}_m^{-1/2}\mathbf{U}\mathbf{D} = \mathbf{X}\mathbf{D}$$

eli

(8) $\mathbf{D}_m^{-1}\mathbf{F}\mathbf{Y} = \mathbf{X}\mathbf{D}$

ja vastaavasti

(9) $\mathbf{D}_n^{1}\mathbf{F}'\mathbf{X} = \mathbf{Y}\mathbf{D}$.

Viimeiset kaksi yhtälöä osoittavat, miten X- ja Y-asteikot liittyvät toisiinsa. Tämä tulee vielä selvemmäksi kirjoittamalla ko. yhtälöt komponenteittain muodossa

$$d_k X_k^{(i)} = \sum_{j=1}^{n} \frac{f_{ij}}{f_{i.}} Y_j^{(k)} , \; i = 1,2,...,m ,$$

$$d_k Y_k^{(j)} = \sum_{i=1}^{m} \frac{f_{ij}}{f_{.j}} X_i^{(k)} , \; j = 1,2,...,n .$$

Asteikkoarvot ovat siis toistensa painotettuja keskiarvoja jaettuina singulaariarvoilla. Tästä johtuu menetelmän eräs varhaisempi nimitys "Method of reciprocal averages". On myös käytetty nimitystä "Dual scaling". Näiden yhtälöiden avulla ratkaisu voidaan laskea iteratiivisesti. Tehokkaampaa on kuitenkin muodostaa tulokset suoraan singulaariarvohajotelman (5) kautta.

Frekvenssimatriisin **F** esityksestä (7) saadaan yhtälöiden (6) avulla

$$(10) \quad \mathbf{F} = \frac{1}{N}(\mathbf{D}_m\mathbf{X})\mathbf{D}(\mathbf{D}_n\mathbf{Y})' = \frac{1}{N}\sum_{k=1}^{n} d_k(\mathbf{D}_m\mathbf{x}^{(k)})(\mathbf{D}_n\mathbf{y}^{(k)})' ,$$

joka on frekvenssimatriisin hajotelma 1-asteisten matriisien summaksi. Näin singulaariarvot d_k , k=1,2,...,n eli maksimaaliset korrelaatiokertoimet ja niitä vastaavat asteikot $\mathbf{x}^{(k)}$ ja $\mathbf{y}^{(k)}$ selittävät voimakkuusjärjestyksessä kaikki frekvenssit. Ensimmäinen dimensio (k=1) vastaa pelkkää matriisin keskistystä ja usein pari seuraavaa dimensioparia (k=2,3) selittää tyydyttävästi loput koko frekvenssitaulukkoon sisältyvästä vaihtelusta.

Korrespondenssianalyysin tulosta havainnollistetaan tavallisesti kuvalla, jossa dimensiot 2 ja 3 asetetaan vastakkain ja sekä rivit että sarakkeet esitetään pisteinä tässä 2-ulotteisessa kuviossa siten, että koordinaatteina ovat ao. asteikkoarvot. Tässä kuvassa kuten yleensäkin korrespondenssianalyysin tulostuksessa asteikot painotetaan vastaavilla singulaariarvoilla eli käytetään asteikkoarvoja **XD** ja **YD** .

Survossa korrespondenssianalyysin toteuttaa operaatio `CORRESP`. Se antaa edellä mainittujen tulosten lisäksi monia muita johdannaisia, joita esitellään korrespondenssianalyysia koskevassa kirjallisuudessa. Päälähteenä on käytetty `CORRESP`-modulia laadittaessa teosta *Lebart*, *Morineau*, *Warwick*: Multivariate Descriptive Statistical Analysis (1984).

10.1 Esimerkki

Käsittelemme Fisherin esimerkkiä eri tavoin. Ensin teemme laskelmia Survon matriisitulkilla käyttämättä CORRESP-operaatiota.

Tässä on tutkittava frekvenssitaulukko muotoiltuna matriisiksi **F**:

```
  1  1 SURVO 84C EDITOR Mon May 02 13:32:43 1994        C:\M\MEN2\ 200 100 0
  1 *
  2 *MATRIX F
  3 * ///    BLUE  LIGHT MEDIUM DARK
  4 * Fair    326    688   343    98
  5 * Red      38    116    84    48
  6 * Medium  241    584   909   403
  7 * Dark    110    188   412   681
  8 * Black     3      4    26    85
  9 *_
```

Seuraava matriisikäskyjen jono tallettaa matriisin ja tekee laskelmat edellä esitettyjen kaavojen mukaisesti:

```
  1  1 SURVO 84C EDITOR Mon May 02 13:35:51 1994        C:\M\MEN2\ 200 100 0
  9 *
 10 *MAT SAVE F
 11 *MAT DN=SUM(F)       / *DN~SUM(F) 1*4
 12 *MAT FT=F'           / *FT~F' 4*5
 13 *MAT DM=SUM(FT)      / *DM~SUM(F') 1*5
 14 *MAT DN2=DN'         / *DN2~SUM(F)' 4*1
 15 *MAT N=SUM(DN2)      / *N~SUM(SUM(F)') D1*1
 16 *MAT TRANSFORM DN2 BY 1/sqrt(X#)
 17 *MAT DN2!=DV(DN2)    / *DN2~DV(T(DN2_by_1/sqrt(X#))) D4*4
 18 *MAT DM2=DM          / *DM2~SUM(F') 1*5
 19 *MAT TRANSFORM DM2 BY 1/sqrt(X#)
 20 *MAT DM2!=DV(DM2)    / *DM2~DV(T(DM2_by_1/sqrt(X#))) D5*5
 21 *MAT DM!=DV(DM)      / *DM~DV(SUM(F')) D5*5
 22 *MAT DN!=DV(DN)      / *DN~DV(SUM(F)) D4*4
 23 *MAT A=DM2*F         / *A~DM2*F 5*4
 24 *MAT A=A*DN2         / *A~DM2*F*DN2 5*4
 25 *MAT SINGULAR_VALUE DECOMPOSITION OF A TO U,D,V
 26 *MAT SN=N            / *SN~SUM(SUM(F)') D1*1
 27 *MAT TRANSFORM SN BY sqrt(X#)
 28 *MAT X=DM2*U         / *X~DM2*Usvd(DM2*F*DN2) 5*4
 29 *MAT X!=SN*X         / *X~T(SN_by_sqrt(X#))*DM2*Usvd(DM2*F*DN2) 5*4
 30 *MAT Y=DN2*V         / *Y~DN2*Vsvd(DM2*F*DN2) 4*4
 31 *MAT Y!=SN*Y         / *Y~T(SN_by_sqrt(X#))*DN2*Vsvd(DM2*F*DN2) 4*4
 32 *_
```

Matriisitiedostojen nimet vastaavat merkintöjä näin:

DM	$\mathbf{D}_m$
DM2	$\mathbf{D}_m^{-1/2}$
DN	$\mathbf{D}_n$
DN2	$\mathbf{D}_n^{1/2}$
SN	$\sqrt{N}$

Muut nimet ovat samoja kuin aikaisemmassa tekstissä. Tärkeimmät tulokset poimittuina Survon toimituskenttään ovat

```
 17  1 SURVO 84C EDITOR Mon May 02 13:53:10 1994        C:\M\MEN2\ 200 100 0
 33 *
 34 *MAT LOAD D,CUR+1
 35 *MATRIX D
 36 *Dsvd(DM2*F*DN2)
 37 *///      sing.val
 38 *svd1     1.000000
 39 *svd2     0.446368
 40 *svd3     0.173455
 41 *svd4     0.029317
 42 *
 43 *MAT LOAD X,CUR+1
 44 *MATRIX X
 45 *///          svd1     svd2     svd3     svd4
 46 *Fair     -1.00000  1.21871 -1.00224 -0.42713
 47 *Red      -1.00000  0.52258 -0.27834  4.02685
 48 *Medium   -1.00000  0.09415  1.20091 -0.11040
 49 *Dark     -1.00000 -1.31888 -0.59929 -0.34507
 50 *Black    -1.00000 -2.45176 -1.65136  1.57370
 51 *
 52 *MAT LOAD Y,CUR+1_
 53 *MATRIX Y
 54 *///          svd1     svd2     svd3     svd4
 55 *BLUE     -1.00000  0.89679 -0.95362 -2.18841
 56 *LIGHT    -1.00000  0.98732 -0.51000  1.08379
 57 *MEDIUM   -1.00000 -0.07531  1.41248 -0.18941
 58 *DARK     -1.00000 -1.57435 -0.77204  0.14822
 59 *
```

ja ne vastaavat (2. pystyrivin osalta) täsmälleen alkuperäisessä taulukossa esitettyjä arvoja. Maksimaalinen korrelaatiokerroin on toinen singulaariarvo 0.446368 rivillä 39.

Samat tulokset voidaan laskea myös iteratiivisesti yhtälöiden (8) ja (9) mukaan, mikä lienee ollut Fisherin alkuperäinen laskentatapa ja siis eräs periaatteessa yksinkertainen keino singulaariarvohajotelman muodostamiseen. Tässä tapauksessa on ensin poistettava ensimmäisen (turhan) singulaariarvon osuus keskistämällä frekvenssimatriisi $\mathbf{F}$ muotoon

$$\mathbf{H} = \mathbf{F} - \mathbf{G} = \mathbf{F} - (\mathbf{D}_m\mathbf{1}_m)(\mathbf{D}_n\mathbf{1}_n)'/N\,.$$

Tässä tapauksessa suoritetaan seuraavat matriisikäskyt:

```
  1  1 SURVO 84C EDITOR Mon May 02 14:14:42 1994        C:\M\MEN2\ 100 100 0
  9 *
 10 *MAT SAVE F
 11 *MAT DN=SUM(F)       / *DN~SUM(F) 1*4
 12 *MAT FT=F'           / *FT~F' 4*5
 13 *MAT DM=SUM(FT)      / *DM~SUM(F') 1*5
 14 *MAT DM2=DM          / *DM2~SUM(F') 1*5
 15 *MAT DN2=DN          / *DN2~SUM(F) 1*4
 16 *MAT TRANSFORM DM2 BY 1/X#
 17 *MAT TRANSFORM DN2 BY 1/X#
 18 *MAT G=MTM2(DM,DN)   / *G~SUM(F')'*SUM(F) 5*4
 19 *MAT N=DM'           / *N~SUM(F')' 5*1
 20 *MAT N=SUM(N)        / *N~SUM(SUM(F')') D1*1
 21 *MAT TRANSFORM N BY 1/X#
 22 *MAT G!=N*G          / *G~T(N_by_1/X#)*SUM(F')'*SUM(F) 5*4
 23 *MAT H=F-G           / *H~F-G 5*4
 24 *MAT DM2!=DV(DM2)      / *DM2~DV(T(DM2_by_1/X#)) D5*5
 25 *MAT DN2!=DV(DN2)      / *DN2~DV(T(DN2_by_1/X#)) D4*4
 26 *MAT DMF=DM2*H       / *DMF~DM2*(F-G) 5*4
 27 *MAT DNFT=MMT2(DN2,H)          / *DNFT~DN2*(F-G)' 4*5
 28 *_
```

Tavoitteena on ollut laskea yhtälöiden (8) ja (9) kerroinmatriisit, joista matriisikäskyissä käytetään merkintöjä

DMF $\mathbf{D}_m^{-1}\mathbf{F}$

DNFT $\mathbf{D}_n^{-1}\mathbf{F}'$.

Iterointi tapahtuu tämän jälkeen seuraavasti:

```
  1  1 SURVO 84C EDITOR Mon May 02 14:26:39 1994        C:\M\MEN2\ 100 100 0
 28 *
 29 *MAT Y=IDN(4,1)
 30 *
 31 *MAT X!=DMF*Y       / *X~DM2*(F-G)*IDN 5*1
 32 *MAT X!=NRM(X)      / *X~NRM(X) 5*1
 33 *MAT Y!=DNFT*X      / *Y~DN2*(F-G)'*X 4*1
 34 *MAT XT=X'          / *XT~X' 1*5
 35 *MAT LOAD XT,END+1
 36 *
 37 *             Fair      Red   Medium     Dark    Black
 38 *  1       0.59553 -0.00273 -0.13455 -0.35562 -0.70765
 39 *  2       0.42496  0.17812  0.00640 -0.42486 -0.77919
 40 *  3       0.39950  0.17084  0.02694 -0.42718 -0.79247
 41 *  4       0.39571  0.16961  0.02998 -0.42746 -0.79438
 42 *  5       0.39513  0.16942  0.03044 -0.42750 -0.79466
 43 *  6       0.39505  0.16939  0.03050 -0.42750 -0.79470
 44 *  7       0.39504  0.16939  0.03051 -0.42750 -0.79471
 45 *  8       0.39503  0.16939  0.03052 -0.42750 -0.79471
 46 *
 47 *MAT LOAD NORM,CUR+1
 48 *MATRIX NORM
 49 *NORM(X)
 50 *///             1
 51 *Norm     0.199245
 52 *
 53 *sqrt(0.199245)=0.44636868169709
 54 *_
```

Rivillä 29 on Y-vektorille annettu alkuarvoksi (1,0,0,0). Riveillä 31-35 on yhtä iteraatiota varten tarvittavat käskyt. Riveillä 38-45 näkyvät 8 ensimmäisen iteraation tuottamat tulokset (tiivistettyinä) X-vektorin osalta. Koska tulos ei ole tässä vaiheessa enää muuttunut, iterointi on lopetettu.

Kullakin kierroksella X-vektori on normeerattu yksikkövektorin mittaiseksi (rivi 32), jolloin `MAT NRM` -komento antaa sivutuloksena alkioitten neliösumman neliöjuuren matriisina `NORM` . Koska iteroinnissa kumpaakin yhtälöistä (8) ja (9) sovelletaan kerran, X-vektori tulee kerrotuksi singulaariarvolla d_2 kahdesti. Tällöin ko. normin neliöjuuri on sama kuin d_2, mikä tulee vahvistetuksi rivillä 53. Vektorin X normeeraus on tässä erilainen kuin ensimmäisessä laskutavassa. Alkiot ovat kuitenkin verrannollisia aikaisempiin.

Jos tällä menettelyllä haluttaisiin määrätä lisää singulaariarvoja ja asteikkoja, frekvenssitaulukosta vähennetään aina edellisen asteikkoparin osuus yhtälön (10) mukaisesti ennen uusia iteraatioita.

Teemme nyt analyysin CORRESP-operaatiolla, joka olettaa, että frekvenssitaulukko **F** on annettu Survon havaintotaulukkona toimituskentässä tai havaintotiedostona. Tässä tapauksessa alkutilanne voisi näyttää seuraavalta:

```
 21  1 SURVO 84C EDITOR Mon May 02 17:35:35 1994        C:\M\MEN2\ 100 100 0
  1 *
  2 *DATA COLORS,A,B,N,M
  3 M ------  AAA    AAA   AAA   AAA  CC.CCC CC.CCC  rrr rrr rrr rrr
  4 N Color  BLUE  LIGHT MEDIUM DARK     C1     C2    BL  LI  ME  DA
  5 A Fair    326    688   343    98
  6 * Red      38    116    84    48
  7 * Medium  241    584   909   403
  8 * Dark    110    188   412   681
  9 B Black     3      4    26    85
 10 *
 11 *CORRESP COLORS,CUR+1_
 12 *
```

Frekvenssitaulukon sarakkeet osoitetaan A-kirjaimella aktivoituina muuttujina ja taulukon rivit (tässä kaikki) aktiivisina havaintoina. Rivikohtaisia tuloksia varten aktivoidaan lisämuuttujia erilaisin kirjaimin. Tässä C tarkoittaa asteikkomuuttujia (X), joita on siis valittu 2. Kirjaimella r aktivoidut muuttujat, joita tulee olla sama määrä kuin taulukossa on sarakkeita, on tarkoitettu residuaalifrekvenssien talletukseen eli siihen osaan matriisin **F** hajotelmasta (10), jota ensimmäiset kaksi dimensiota tässä tapauksessa eivät selitä.

Muut tulostusmahdollisuudet ilmenevät CORRESP-operaation kuvauksesta Survon neuvontajärjestelmässä ja em. Lebartin, Morineaun ja Warwickin kirjasta.

Kun CORRESP aktivoidaan, saadaan tulokset:

```
 21  1 SURVO 84C EDITOR Mon May 02 17:50:43 1994        C:\M\MEN2\ 100 100 0
  1 *
  2 *DATA COLORS,A,B,N,M
  3 M ------  AAA    AAA   AAA   AAA  CC.CCC CC.CCC  rrr rrr rrr rrr
  4 N Color  BLUE  LIGHT MEDIUM DARK     C1     C2    BL  LI  ME  DA
  5 A Fair    326    688   343    98 -0.544 -0.174     5  -6   1  -1
  6 * Red      38    116    84    48 -0.233 -0.048   -10  11  -2   1
  7 * Medium  241    584   909   403 -0.042  0.208     2  -2   0  -0
  8 * Dark    110    188   412   681  0.589 -0.104     4  -4   1  -1
  9 B Black     3      4    26    85  1.094 -0.286    -2   2  -0   0
 10 *
 11 *CORRESP COLORS,CUR+1_
 12 *Correspondence analysis on data COLORS: Rows=5 Columns=4
 13 *
 14 *    Canonical      Eigen-           Chi^2    Cumulative
 15 *    correlation    value                     percentage
 16 * 1  0.4464         0.1992      1073.33148       86.56
 17 * 2  0.1735         0.0301      162.077452       99.63
 18 * 3  0.0293         0.0009      4.63002608      100.00
 19 *                   0.2302     1240.04 (df=12 P=0)
 20 *
 21 *Column coordinates (CR_COORD.M)
 22 *             C1     C2
 23 *BLUE     -0.400 -0.165
 24 *LIGHT    -0.441 -0.088
 25 *MEDIUM    0.034  0.245
 26 *DARK      0.703 -0.134
 27 *
```

Varatut lisäsarakkeet ovat täyttyneet rivikohtaisilla tuloksilla. Vastaavat sarakekohtaiset tulokset talletetaan matriisitiedostoiksi. Tässä tapauksessa sarake-

koordinaatit ovat matriisitiedostossa CR_COORD.M, joka tulostuu yhteenvetotaulukon perään riveille 21-26.

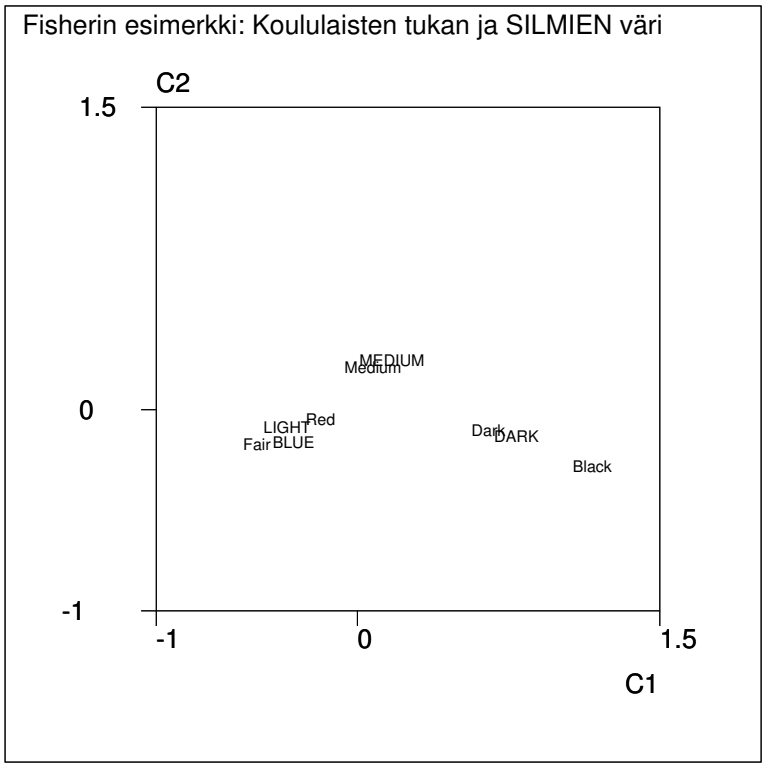
Yhteenvetotaulukossa riveillä 14-19 annetaan maksimikorrelaatiot d_k, k=2,3,4 ja toisena sarakkeena näiden neliöt (matriisin **A**'**A** ominaisarvot), jotka usein tulkitaan selitysosuuksiksi, vaikka frekvenssien tasolla pikemminkin singulaariarvoilla on tämä luonne yhtälön (10) mukaan.

Chi^2-sarakkeessa ovat luvut Nd_k^2. Näiden summa on itse asiassa tavanomainen χ^2-testisuure, joka noudattaa likimain χ^2-jakaumaa (m-1)(n-1) vapausasteella, kun taulukoitujen muuttujien välillä ei ole lainkaan riippuvuutta.

Jäännösfrekvensseistä (rrr-sarakkeet) useimmat ovat lähellä nollaa. Jos punatukkaisista ja vaaleasilmäisistä siirrettäisiin 10 sinisilmäisten luokkaan sekä vaaleatukkaisista ja sinisilmäisistä 5 vaaleasilmäisiin, jäännökset tulisivat tasaisesti hyvin pieniksi. Tämän voi tarkistaa tekemällä ao. muutokset frekvenssitaulukossa ja toistamalla analyysin. Muutos ei kuitenkaan vaikuta juuri lainkaan muihin tuloksiin.

Korrespondenssianalyysille ominainen kaksiulotteinen esitys, jossa sekä rivit että sarakkeet esitetään päällekkäin edellä saatujen koordinaattien osoittamissa paikoissa, luodaan kuvaruutuun kahdella GPLOT-kaaviolla:

```
 23  1 SURVO 84C EDITOR Mon May 02 18:47:34 1994        C:\M\MEN2\ 100 100 0
 28 *.....................................................................
 29 *HEADER=Fisherin_esimerkki:_Koululaisten_tukan_ja_silmien_väri
 30 *GPLOT COLORS,C1,C2 /     POINT=[RED],Color OUTFILE=A
 31 *SCALE=-1,0,1.5 MODE=VGA XDIV=179,380,80 YDIV=59,380,40
 32 *.....................................................................
 33 *GPLOT CR_COORD.M,C1,C2_/ POINT=[BLUE],CASE HEADER= INFILE=A
 34 *SCALE=-1,0,1.5 MODE=VGA XDIV=179,380,80 YDIV=59,380,40
 35 *
```



Tämä kuvaustapa on sinänsä mielenkiintoinen ja antaa tässä sovelluksessa hyvin uskottavan käsityksen ao. ominaisuuksien suhteesta. Kun kuitenkin muistetaan, että tukan ja silmien väliset maksimaaliset korrelaatiot olivat vain 0.45 ja 0.17, rohkeita tulkintoja tulee välttää.

Moniulotteisista kuutioista ja palloista

Moniulotteisen avaruuden hahmottaminen ei ole helppoa. Tästä huolimatta käytämme kuitenkin mielellämme 2- ja 3-ulotteisia mielikuviamme malleina useampiulotteisissa tilanteissa. Nämä mielikuvat toimivat hyvin useissa tapauksissa, mutta joskus ne voivat olla aika harhaanjohtavia. On hyvää harjoitusta yrittää venyttää omaa mielikuvitustaan ja intuitiivisesti yleistää alempiulotteisia ominaisuuksia samalla varoen tekemästä liian rohkeita päätelmiä.

Tarkastelkaamme yksinkertaisena esimerkkinä kuutiota ja palloa ja miettikäämme, mitä ominaisuuksia on niiden yleistyksillä n-ulotteisessa avaruudessa.

Kuution osat

Aloitamme kuutiosta ja yritämme osaksi arvaillen pohtia, millainen olisi sen yleistys 4-, 5- jne. -ulotteisessa avaruudessa. Kuutio koostuu tunnetusti 8 kärkipisteestä, joita yhdistää 12 särmää. Kuutiolla on edelleen nimensä mukaisesti 6 neliömäistä sivutahkoa. Neliötä voi pitää 2-ulotteisena kuutiona, särmää 1-ulotteisena kuutiona ja (kärki)pistettä 0-ulotteisena kuutiona. Näin ollen kuutio koostuu alempiulotteisista "kuutioista" seuraavan taulukon mukaisesti:

n/m	0	1	2	3	4	summa
0	1					1
1	2	1				3
2	4	4	1			9
3	8	12	6	1		27
4	16	x	y	8	1	z

Taulukosta näkyy kuinka monesta m-ulotteisesta kuutiosta koostuu n-ulotteinen kuutio. Merkitsemme tätä lukumäärää yleisesti $K(n,m)$. On luonnollista asettaa $K(n,n)=1$. Näin arvoilla $n=0,1,2,3$ taulukossa esitetyt luvut ovat kiistattomia.

Sen sijaan arvolla $n=4$ on toistaiseksi esitettävä vain arvauksia. On ilmeistä, että dimension kasvaessa yhdellä kärkipisteiden määrä kaksinkertaistuu. Näin ollen $K(n,0)=2^n$ ja siis $K(4,0)=16$. Samoin taulukon perusteella on arvattavissa, että $K(n,n\text{-}1)=2n$ eli $K(4,3)=8$.

Taulukkoon lasketusta summasarakkeesta voi päätellä, että ilmeisesti n-ulotteisen kuution tällä tavalla määriteltyjen komponenttien yhteenlaskettu lukumäärä on 3^n eli $z=81$. Tuntemattomiksi jäävät enää luvut $x=K(4,1)$ ja $y=K(4,2)$ joiden summa on $x+y=81\text{-}16\text{-}8\text{-}1=56$. Todetaan, että $K(n,1)=K(n,0)n/2$, sillä kustakin kärkipisteestä lähtee n särmää. Siis $x=K(4,1)=16*4/2=32$ ja $y=K(4,2)=56\text{-}x=24$.

Näillä arvauksilla täydennettynä taulukko näyttää seuraavalta:

n/*m*	0	1	2	3	4	summa
0	1					1
1	2	1				3
2	4	4	1			9
3	8	12	6	1		27
4	16	32	24	8	1	81

Jos tästä yrittää saada vielä yleisempiä päätelmiä, on ratkaisevaa havaita, että se muistuttaa (rivit 0 ja 1) jossain määrin Pascalin kolmiota. Ei liene aivan mahdotonta huomata tällöin, että luvulle $K(n,m)$ pätee palautuskaava

$$K(n,m)=K(n\text{-}1,m\text{-}1)+2K(n\text{-}1,m)$$

eli siis ainoastaan kerroin 2 erottaa tämän binomikertoimien palautuskaavasta. On tämän jälkeen helppo havaita, että jakamalla taulukon luvut luvuilla 2^{n-m} ne muuntuvat binomikertoimiksi $C(n,m)$ eli siis

$$K(n,m)=C(n,m)2^{n-m} .$$

Huom. Jos samalla tavalla analysoi n-ulotteisen simpleksin rakennetta (1-ulotteinen on jana, 2-ulotteinen on kolmio, 3-ulotteinen on nelitahokas), päädytään suoraan Pascalin kolmioon.

Tämän esityksen perusteella taulukko täydentyisi seuraavasti:

n/*m*	0	1	2	3	4	5	6	7	8	summa
0	1									1
1	2	1								3
2	4	4	1							9
3	8	12	6	1						27
4	16	32	24	8	1					81
5	32	80	80	40	10	1				243
6	64	192	240	160	60	12	1			729
7	128	448	672	560	280	84	14	1		2187
8	256	1024	1792	1792	1120	448	112	16	1	6561

Näemme, että esim. 8-ulotteisen kuution täytyy olla jo melko konstikas "kappale".

Tehdäksemme päättelymme pitävämmäksi on parasta siirtyä analyyttisempaan tarkastelutapaan samaistamalla kuution n-ulotteisen avaruuden pisteeseen (1,1,...,1) ja tämän pisteen projektiopisteisiin koordinaattiakseleille, -tasoille jne. Kuutiota rajoittavat kärkipisteet ovat siis $(e_1,e_2,...,e_n)$, missä jokainen e_i

on joko 0 tai 1.

Näin määritellyn n-ulotteisen kuution komponentteja (alempiulotteisia osakuutioita) voidaan merkitä seuraavasti. Esim. kolmiulotteisessa tapauksessa (x,0,0) tarkoittaa sitä osaa, joka syntyy kun x vaihtelee välillä (0,1) eli särmää joka yhtyy ensimmäiseen koordinaattiakseliin. (x,x,0) taas merkitsee neliötä, joka on kahden ensimmäisen koordinaattiakselin muodostamassa tasossa ja (x,x,1) edellisen neliön vastakkaista sivutahkoa kuutiossa. (x,x,x) on koko kuutio.

n-ulotteisen kuution m-ulotteisen osakuution merkinnässä on tällöin m kappaletta x:iä ja loput koordinaatit ovat nollia tai ykkösiä. Koska "vapaiden" koordinaattien x paikat voidaan valita binomikertoimen $C(n,m)$ osoittamalla tavalla ja loput n-m paikkaa merkitä joko 0 tai 1 tästä riippumatta 2^{n-m} tavalla, on

$$K(n,m)=C(n,m)2^{n-m} ,$$

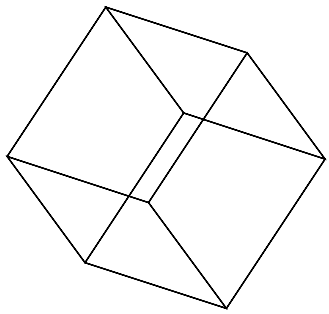
mikä vahvistaa aikaisemman arvauksen.

Lukujen $K(n,m)$, m=0,1,...,n summaksi todennetaan 3^n yksinkertaisesti soveltamalla binomikaavaa potenssiin $(1+2)^n$.

Jo tähänastinen tarkastelu osoittaa, miten analyyttinen ote puree paremmin kuin geometriset mielikuvat.

Neliulotteisen kuution 2-ulotteinen projektio

Kun piirrämme kuvan kuutiosta tasoon, se merkitsee kuution (tavallisesti yhdensuuntaista) projisointia tasoon ja kuva näyttää seuraavalta:



Tutkikaamme vastaavasti, miltä 4-ulotteinen kuutio näyttäisi tasolle projisoituna.

Toteutamme tämän Survon avulla. Rakennamme aluksi 33 x 4 matriisin, jonka riveinä ovat 4-ulotteisen kuution kärkipisteet sellaisessa järjestyksessä, että yhdistämällä peräkkäiset pisteet kaikki 32 särmää tulevat läpikäydyiksi ja piirretyiksi. Tällaista reittiä, joka käy läpi kaikki verkon särmät ja päätyy al-

kupisteeseen sanotaan Hamiltonin reitiksi. Sellaisia löytyy yhtenäisistä verkoista, joiden kaikkien pisteiden asteluvut ovat parilliset. Neliulotteisessa kuutiossa jokaiseen kärkipisteeseen liittyy 4 särmää ja asteluku on siis 4. Reitti kulkee kahdesti kunkin kärjen kautta. Tavallisessa kuutiossa Hamiltonin reittiä ei ole (kärkien asteluvut =3), mutta neliössä (asteluvut=2) se saadaan yksinkertaisesti kiertämällä ympäri.

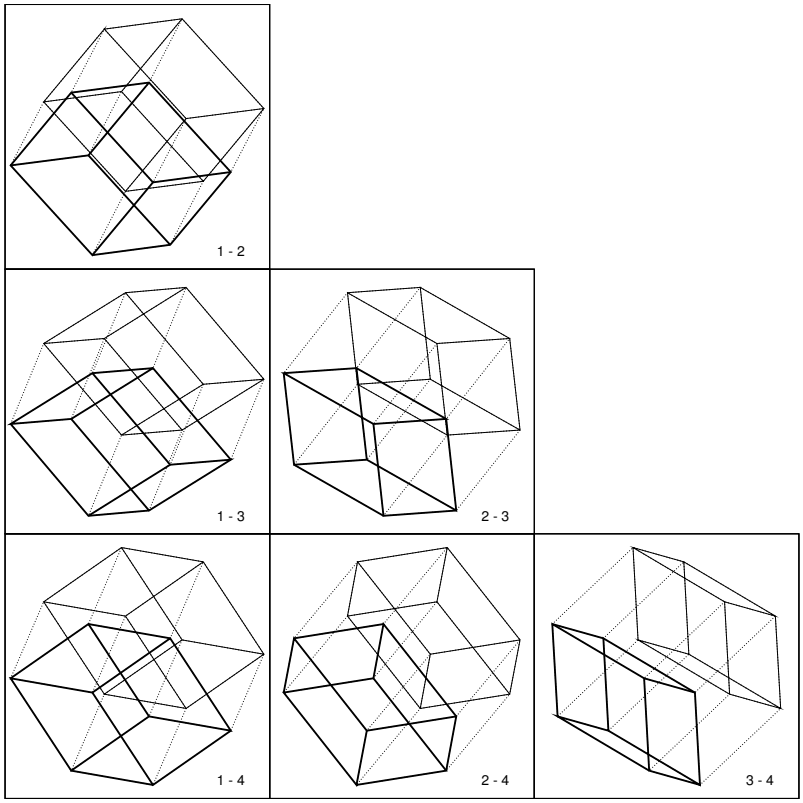
```
  8  1 SURVO 84C EDITOR Mon May 25 10:59:39 1993       D:\MON\DIM\ 120  80 0
  1 *
  2 *Hamiltonin reitti 4-ulotteisessa kuutiossa:
  3 *MATRIX H ///
  4 *0 0 0 0
  5 *0 0 0 1
  6 *0 0 1 1
  7 *0 0 1 0
  8 *0 0 0 0
  9 *0 1 0 0
 10 *0 1 0 1
 11 *0 1 1 1
 12 *0 1 1 0
 13 *0 1 0 0
 14 *1 1 0 0
 15 *1 1 0 1
 16 *0 1 0 1
 17 *0 0 0 1
 18 *1 0 0 1
 19 *1 1 0 1
 20 *1 1 1 1
 21 *0 1 1 1
 22 *0 0 1 1
 23 *1 0 1 1
 24 *1 0 0 1
 25 *1 0 0 0
 26 *1 1 0 0
 27 *1 1 1 0
 28 *0 1 1 0
 29 *0 0 1 0
 30 *1 0 1 0
 31 *1 0 1 1
 32 *1 1 1 1
 33 *1 1 1 0
 34 *1 0 1 0
 35 *1 0 0 0
 36 *0 0 0 0
 37 *
```

Talletamme matriisin **H**, siirrämme keskipisteen origoon ja kierrämme kuviota mielivaltaisesti kertomalla matriisilla **H** ortogonaalisen 4 x 4 -matriisin **Q**. Näin syntyy kierretyn neliulotteisen kuution kärkiä kuvaava matriisi **K=HQ**. Matriisi **Q** on muodostettu tekemällä "mielivaltaiselle" matriisille **T** Gram-Schmidt-hajotelma **T=QR**, missä **Q** on ortogonaalinen ja **R** kolmiomatriisi. Kaikki tämä saadaan Survossa aikaan seuraavasti:

```
 23  1 SURVO 84C EDITOR Tue May 25 11:07:16 1993        D:\MON\DIM\ 120  80 0
 37 *
 38 *MAT SAVE H
 39 *MAT TRANSFORM H BY X#-0.5   / Keskistys  (0,1) -> (-0.5,0.5)
 40 *MAT CLABELS "X" TO H        / Sarakeotsikot X1,X2,X3,X4
 41 *
 42 *MAT T=ZER(4,4)
 43 *MAT TRANSFORM T BY sin(31*I#*J#)  / "Mielivaltainen" T
 44 *
 45 *MAT GRAM-SCHMIDT DECOMPOSITION OF T TO Q,R  / T:n ortogonalisointi
 46 *MAT LOAD Q,CUR+1
 47 *MATRIX Q
 48 *GS(T(T_by_sin(31*I#*J#)))
 49 *///            1         2         3         4
 50 *  1       -0.25056 -0.48181 -0.60885 -0.57825
 51 *  2       -0.45840 -0.54993 -0.03826  0.69712
 52 *  3       -0.58807 -0.07771  0.69287 -0.40997
 53 *  4       -0.61747  0.67778 -0.38441  0.10756
 54 *
 55 *MAT K=H*Q                  / Neliulotteisen kuution kierto
 56 *MAT CLABELS "dim" TO K_    / Sarakeotsikot dim1,dim2,dim3,dim4
 57 *
```

Kuvion kaksiulotteiset projektiot koordinaattitasoille saadaan tavallisena XY-piirroksena "unohtamalla" muut kaksi dimensiota (pystyriviä). Särmät tulevat näkyviin yhdistämällä peräkkäiset pisteet suorilla viivoilla. Ilman kiertoa kaikissa projektioissa näkyisi pelkkä neliö.

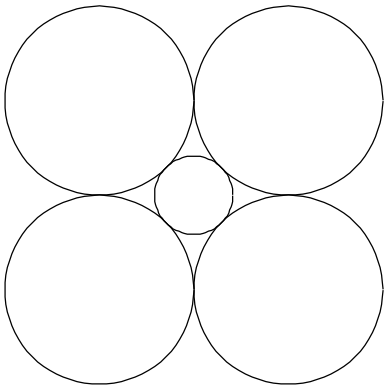
Seuraava kuva esittää kaikki mahdolliset projektiot yhtenä kuvamatriisina. Tiettyjen vastinkuutioiden hahmottamiseksi kuvassa on käytetty 3 eri viivatyyppiä. Kaikki tähän esimerkkiin liittyvät tarkemmat tiedot löytyvät toimituskentästä KUUTIO4.



Pallot n-ulotteisina

Aloitamme paradoksaaliselta kuulostavalla esimerkillä, jonka kuulin prof. Shirjayevilta ja jota hänen kertomansa mukaan jo esim. A.N. Kolmogorov (todennäköisyyslaskennan venäläinen uranuurtaja) oli käyttänyt luennoillaan.

Tarkastellaan neljää yksikkösäteistä ympyrää, jotka sivuavat toisiaan oheisen kuvan mukaisesti ja joiden keskelle on piirretty kaikkia neljää sivuava (paljon pienempi) ympyrä:



On helppo todeta, että pienen ympyrän säde on $\sqrt{2}-1 \approx 0.4142$. Kun vastaava konstruktio toteutetaan kolmiulotteisesti asettamalla 8 yksikkösäteistä palloa kasaan, jää keskelle tilaa pienemmälle pallolle, joka sivuaa kaikkia kahdeksaa. Tämän pallon säteeksi saadaan $\sqrt{3}-1 \approx 0.7321$. Yleisesti on mahdollista päätellä, että vastaavassa n-ulotteisessa rakennelmassa pienen n-ulotteisen pallon säteeksi saataisiin $\sqrt{n}-1$. Siis esim. kun n=4, "pieni" pallo olisi samankokoinen kuin sitä ympäröivät ja arvolla n=9 sen säde olisi kaksinkertainen. Dimension n kasvaessa keskellä olevan pallon säde kasvaisi rajatta. Missä on vika?

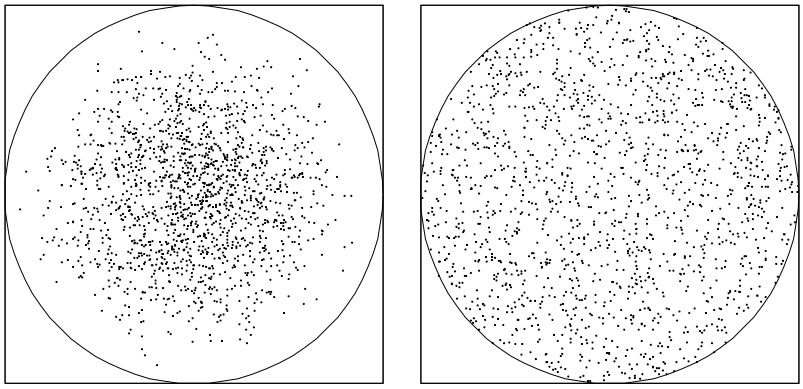
Vikaa ei ole missään. Käsityksemme n-ulotteisen pallon olemuksesta vain on harhaanjohtava. Se on paljon "hennompi olento" kuin mitä saatamme kuvitella 2- ja 3-ulotteisten mielikuvien varassa. Ajatelkaamme n-ulotteista kuutiota, jonka sivun pituus on 2 ja sen sisälle asetettua mahdollisimman suurta n-ulotteista palloa, jonka säde on siis 1. Pallon suurin ulottuvuus on sen halkaisija =2, mutta kuution suurin ulottuvuus on sen päälävistäjä, jonka pituus on $2\sqrt{n}$ (saadaan soveltamalla Pythagoraan lausetta n-1 kertaa). Pallo täyttää kuution hyvin vajaasti ja näin jää toisiaan sivuavien pallojen väliin tilaa paljon enemmän kuin saatamme etukäteen kuvitella.

Kun $n=2k$, n-ulotteisen, R-säteisen pallon mitta (tilavuus) on $\pi^k R^{2k}/k!$ (johdetaan myöhemmin) ja sen suhde ympäröivän kuution mittaan $(2R)^{2k}$ lähestyy sangen pikaisesti nollaa, kuten nähdään taulukosta:

n	suhde
2	0.78540
4	0.30843
6	0.08075
8	0.01585
10	0.00249
12	0.00033

Seuraava kuvapari näyttää, miten pallon sisältö kasautuu sitä enemmän keskipisteen läheisyyteen, mitä suuremmaksi ulotteisuus n kasvaa. Olemme arponeet tasajakaumaperiaatteella 8-ulotteisen kuution sisältä 100000 pistettä ja valinneet ne, jotka ovat myös tämän kuution sisään asetetun 8-ulotteisen pallon sisällä. Näitä pisteitä löytyi 1635 kappaletta (odotettu arvo olisi taulukon mukaan 1585 keskivirheellä 39.5) eli uskottava määrä.

Jos nyt projisoimme tämän 8-ulotteisen palloa hahmottavan pistekuvion tasoon, saamme seuraavanlaisen 2-ulotteisen jakauman. Vertailun vuoksi, viereinen, oikeanpuolinen kuva näyttää neliön sisälle piirrettyyn ympyrään tasajakauman mukaisesti arvotut 1635 pistettä.



8-ulotteisen pallon projektiossa pisteet eivät enää jakaudu tasaisesti ympyrälle, vaan ne keskittyvät voimakkaasti kohti keskustaa. Tämäkin osoittaa, kuinka hintelä on moniulotteinen pallo verrattuna sitä ympäröivään moniulotteiseen kuutioon. Tutkiessamme seuraavaksi tasaista jakaumaa n-ulotteisessa pallossa tulemme osoittamaan, että näiden projektiopisteiden yhteisjakauman tiheysfunktio on tässä tapauksessa $f(x,y)$=vakio$\cdot(1-x^2-y^2)^3$.

Tasainen jakauma n-ulotteisessa pallossa

Todistaaksemme eräät edellämainitut tulokset tarkastelemme tasaista jakaumaa R-säteisessä, n-ulotteisessa, origokeskisessä pallossa, joka määritellään avaruudessa $\mathbf{R}^n$ niiden pisteiden joukkona, joiden etäisyys origosta on korkeintaan R eli

$$(1) \quad \{ \mathbf{x}=(x_1,x_2,...,x_n) \mid x_1^2+x_2^2+ ... +x_n^2 \le R^2 \}$$

Tarvitsemme jatkuvasti seuraavaa muunnosta, joka on tavallisten napa- ja pallokoordinaattimuunnosten yleistys. Muunnoksen välittävät yhtälöt

$$\begin{aligned}
x_1 &= r\cos\varphi_1 \\
x_2 &= r\sin\varphi_1\cos\varphi_2 \\
x_3 &= r\sin\varphi_1\sin\varphi_2\cos\varphi_3 \\
&... \\
x_{n-1} &= r\sin\varphi_1\sin\varphi_2\sin\varphi_3 ... \sin\varphi_{n-2}\cos\varphi_{n-1} \\
x_n &= r\sin\varphi_1\sin\varphi_2\sin\varphi_3 ... \sin\varphi_{n-2}\sin\varphi_{n-1} .
\end{aligned}$$

Tässä r on pisteen $\mathbf{x}$ etäisyys origosta

$$r = \|\mathbf{x}\| = \sqrt{x_1^2+x_2^2+...+x_n^2}$$

ja kulmat $\varphi_1, \varphi_2,..., \varphi_{n-1}$ määräytyvät yhtälöistä

$$\varphi_i = \arctan(\sqrt{x_{i+1}^2+x_{i+2}^2+...+x_n^2} \,/\, x_i), \quad i=1,2,...,n-1 .$$

Tämä muunnos kuvaa pallon (1) n-ulotteiseksi suorakulmioksi

$$\{ r,\varphi_1,...,\varphi_{n-1} \mid 0\le r\le R,\ 0\le\varphi_1\le\pi,...,\ 0\le\varphi_{n-2}\le\pi,\ 0\le\varphi_{n-1}\le 2\pi \}$$

ja sen funktionaalideterminantti on

$$(2) \quad D_n = \frac{\partial(x_1,x_2,...,x_n)}{\partial(r,\varphi_1,...,\varphi_{n-1})} = r^{n-1}\sin^{n-2}\varphi_1\sin^{n-3}\varphi_2 ... \sin^2\varphi_{n-2}\sin\varphi_{n-1} .$$

Tämä osoitetaan muodostamalla ko. determinantin alkiot, ja kehittämällä determinantti viimeisen pystyrivin mukaan (jossa vain kaksi viimeistä alkiota poikkeaa nollasta). Tällöin saadaan

$$\begin{aligned}
D_n &= r\sin\varphi_1\sin\varphi_2 ... \sin\varphi_{n-2}(\cos^2\varphi_{n-2}D_{n-1}+ \sin^2\varphi_{n-2}D_{n-1}) \\
&= r\sin\varphi_1\sin\varphi_2 ... \sin\varphi_{n-2}D_{n-1} .
\end{aligned}$$

Soveltamalla tätä kaavaa n-1 kertaa ja ottamalla huomioon, että $D_1=r$, saadaan D_n:lle lauseke (2).

Laskemme nyt aluksi n-ulotteisen pallon mitan ("tilavuuden") V_n:

$$\begin{aligned} V_n &= \int_{r \le R} dx_1 dx_2 \ldots dx_n \\ &= \int_0^R \int_0^\pi \ldots \int_0^\pi \int_0^{2\pi} D_n \, dr d\varphi_1 \ldots d\varphi_{n-2} d\varphi_{n-1} \\ &= \int_0^R r^{n-1} dr \int_0^\pi \sin^{n-2}\varphi_1 \, d\varphi_1 \int_0^\pi \sin^{n-3}\varphi_2 \, d\varphi_2 \ldots \int_0^\pi \sin\varphi_{n-2} \, d\varphi_{n-2} \int_0^{2\pi} d\varphi_{n-1} \\ &= 2\pi R^n / n \; S_{n-2} S_{n-3} \ldots S_1 , \end{aligned}$$

missä on merkitty

$$S_m = \int_0^\pi \sin^m\varphi \, d\varphi .$$

Osittaisintegroinnilla todetaan, että

$$S_m = \frac{m-1}{m} S_{m-2} \quad \text{ja} \quad S_2 = \pi/2, \; S_1 = 2$$

ja tämän perusteella edelleen, että

$$S_m S_{m-1} = 2\pi/m .$$

Edellä saadusta tilavuuden V_n lausekkeesta ja näistä yhtälöistä seuraa, että

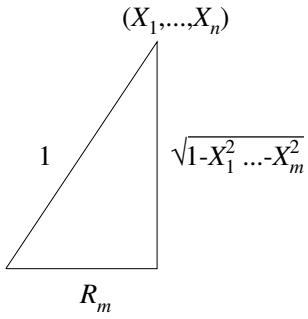
$$V_n = R(n-1)/n S_{n-2} V_{n-1} = R S_n V_{n-1} = R^2 S_n S_{n-1} V_{n-2} = 2\pi/n R^2 V_{n-2} .$$

Koska tunnetusti $V_1 = 2R$ ja $V_2 = \pi R^2$, saamme tästä suoraan yleiset lausekkeet erikseen parillisille ja parittomille dimensioille. Esim. parillisilla arvoilla $n = 2k$ on $V_{2k} = \pi^k R^{2k}/k!$.

Ryhdymme nyt tarkastelemaan tasaista jakaumaa yksikkösäteisessä, n-ulotteisessa pallossa. Tutkimme tämän jakauman m-ulotteista reunajakaumaa (symmetriasta johtuen kaikki samanulotteiset reunajakaumat ovat keskenään identtisiä) ja erityisesti kiinnostaa satunnaisen pisteen $(X_1,\ldots,X_n)$ m-ulotteisen projektion $(X_1,\ldots,X_m)$ origosta lasketun etäisyyden

$$R_m = \sqrt{X_1^2 + \ldots + X_m^2}$$

jakauma.



Esitämme nyt seuraavat väitteet:

$$f_{X_1,...,X_m}(x_1,...,x_m) \propto (1-x_1^2-... -x_m^2)^{\frac{n-m}{2}}, \quad \text{kun } x_1^2+...+x_m^2 \le 1 ,$$

$$f_{R_m}(r) \propto r^{m-1}(1-r^2)^{\frac{n-m}{2}}, \quad 0 \le r \le 1 .$$

Jälkimmäisestä seuraa suoraan, että R_m^2 noudattaa Beta-jakaumaa parametrein $m/2$ ja $(n-m)/2+1$.

Osoitamme väitteet oikeiksi tavanomaisella induktiolla. Koska

$$f_{X_1,...,X_n}(x_1,...,x_n) = \text{vakio}$$

ja

$$f_{R_n}(u) = \int_0^{\pi} \dots \int_0^{\pi} \int_0^{2\pi} D_n \, d\varphi_1 \dots d\varphi_{n-2} d\varphi_{n-1} \propto r^{n-1} ,$$

väitteet pitävät paikkansa, kun $m=n$. Näytämme nyt, että siitä, että ne pitävät paikkansa arvolla m, seuraa, että ne ovat voimassa myös arvolla m-1. Merkitsemällä $u = \sqrt{1-x_1^2-...-x_{m-1}^2}$ saamme

$$f_{X_1,...,X_{m-1}}(x_1,...,x_{m-1}) \propto \int_0^u (u^2-x_m^2)^{\frac{n-m}{2}} dx_m \propto u^{n-m+1} = (1-x_1^2-... -x_m^2)^{\frac{n-(m-1)}{2}} ,$$

mikä on väitettyä muotoa. Siirtymällä m-1-ulotteisella pallokoordinaatistomuunnoksella muuttujiin $r, \varphi_1,...,\varphi_{m-2}$ saamme

$$f_{R_{m-1}}(r) \propto \int_0^{\pi} \int_0^{\pi} \dots \int_0^{2\pi} D_{m-1}(1-r^2)^{\frac{n-(m-1)}{2}} \, d\varphi_1 d\varphi_2 ... d\varphi_{m-2} \propto r^{m-2}(1-r^2)^{\frac{n-(m-1)}{2}}$$

eli sekin on väitteen mukainen. Induktiopäättely osoittaa väitteet oikeaksi kaikilla arvoilla $m=n,n-1,...,2,1$.

Näin on vahvistettu ne toteamukset, jotka esitimme edellisessä luvussa. Kun simuloimme tasaista jakaumaa hylkäysperiaatteella, 8-ulotteisessa tapauksessa vain noin 1.5 % pisteistä osui pallon sisälle. Saatujen tulosten avulla ko. simulointi voidaan toteuttaa huomattavasti tehokkaammin arpomalla suoraan kulmat $\varphi_1,...,\varphi_{n-2}$ tasaisesta jakaumasta välillä $(0,\pi)$, kulma φ_{n-1} tasaisesta jakaumasta välillä $(0,2\pi)$ ja säde r tasaisesti välillä (0,1) jakautuneen muuttujan u

potenssina $r=u^{1/n}$.

Jos halutaan simuloida suoraan m-ulotteista reunajakaumaa, kulmat arvotaan kuten edellä ja säde r neliöjuurena Beta-jakautuneesta satunnaisluvusta parametrein $m/2$, $(n-m)/2+1$. Kun tiedetään, että välin (0,1) tasaisesta jakaumasta saadussa N havainnon otoksessa M:nneksi pienin havainto noudattaa Beta-jakaumaa parametrein M, $N-M+1$, voidaan tätä käyttää hyväksi r-arvojen generoinnissa, kun n ja m ovat parillisia lukuja. Esim. tapauksessa $n=8$, $m=2$, josta edellä oli kuva, r saadaan kolmesta satunnaisluvusta pienimmän neliöjuurena.

Huomaa, ettei edes tapauksessa $n=m=2$ (tasainen jakauma ympyrässä), r ei ole tasaisesti jakautunut, mutta r^2 on.

Singulaariarvo- ja muita hajotelmia matriiseille

Tämä on lyhyt katsaus matriisihajotelmiin, joilla on sovelluksia tilastollisisssa monimuuttujamenetelmissä. Singulaariarvohajotelma (Singular Value Decomposition, lyhennettynä SVD) on merkityksestään huolimatta jäänyt useimmissa alan oppikirjoissa jatkuvasti vaille riittävää huomiota, vaikka sen avulla hyvin monet tulokset ovat kaikkein helpoimmin johdettavissa. Tämä on sitäkin hämmästyttävämpää, kun kuitenkin monet oppikirjat sen mainitsevat, mutta eivät jostain syystä käytä sitä tehokkaasti hyväksi.

SVD:n historia on jonkin verran hämärän peitossa. Kunnolla julkisuuteen se lienee tullut vasta vuonna 1936 *Eckartin* ja *Youngin* artikkelissa Psychometrika-lehdessä ja kulkeekin usein myös Eckart-Young-hajotelman nimellä.

Ennen SVD:n ja siihen perustuvien tärkeiden matriisien ominaisuuksia koskevien tulosten esittelyä käydään lyhyesti läpi eräät muut matriisihajotelmat. Tämä katsaus ei pyri ehdottomaan matemaattiseen tarkkuuteen. Ainoastaan sellaiset asiat pyritään perustelemaan tarkemmin, joita ei välttämättä löydy alan oppikirjoista.

Spektraalihajotelma

Olkoon $\mathbf{A}$ symmetrinen $n\times n$-matriisi. Tällöin

$$\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}',$$

missä Λ on ominaisarvojen $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ muodostama lävistäjämatriisi, ja $\mathbf{U}$ on näitä vastaavien ominaisvektorien $\mathbf{u}^{(1)}$, $\mathbf{u}^{(2)}$,..., $\mathbf{u}^{(n)}$ muodostama ortogonaalinen $n\times n$-matriisi

$$\mathbf{U} = [\ \mathbf{u}^{(1)}\ \mathbf{u}^{(2)}\ ...\ \mathbf{u}^{(n)}\ ]$$

eli $\mathbf{U}'\mathbf{U}=\mathbf{I}$ ja $\mathbf{U}\mathbf{U}'=\mathbf{I}$. Spektraalihajotelma kirjoitetaan usein myös summamuodossa

$$\mathbf{A} = \lambda_1\mathbf{u}^{(1)}\mathbf{u}^{(1)'} + \lambda_2\mathbf{u}^{(2)}\mathbf{u}^{(2)'} + ... + \lambda_n\mathbf{u}^{(n)}\mathbf{u}^{(n)'}.$$

Gram-Schmidt-hajotelma

Olkoon $m\times n$-matriisi $\mathbf{A}$, missä $m\geq n$, täysiasteinen eli $r(\mathbf{A})=n$. Tällöin

$$\mathbf{A} = \mathbf{Q}\mathbf{R},$$

missä $\mathbf{Q}$ on pystyriveittäin ortogonaalinen $m\times n$-matriisi ($\mathbf{Q}'\mathbf{Q}=\mathbf{I}$) ja $\mathbf{R}$ $n\times n$-yläkolmiomatriisi.

Cholesky-hajotelma

Olkoon $\mathbf{A}$ $n\times n$-positiivisesti definiitti matriisi (eli $\mathbf{A}>0$). Tämä tarkoittaa määritelmällisesti sitä, että neliömuoto $\mathbf{u}'\mathbf{A}\mathbf{u} > 0$ kaikilla n komponentin pystyvektoreilla $\mathbf{u}\neq\mathbf{0}$. Tästä on seurauksena mm. että $\mathbf{A}$:n ominaisarvot ja sen determinantti (= ominaisarvojen tulo) ovat positiivisia. Tällöin $\mathbf{A}$:n Cholesky-hajotelma on

$$\mathbf{A} = \mathbf{L}\mathbf{L}',$$

missä $\mathbf{L}$ on ylä- (tai vaihtoehtoisesti ala-) kolmiomatriisi, jonka kaikki lävistäjäalkiot ovat positiivisia.

Singulaariarvohajotelma

Olkoon **A** $m\times n$-matriisi, missä $m\geq n$. Ilman mitään lisäoletuksia **A** voidaan aina esittää muodossa

$$\mathbf{A} = \mathbf{UDV}' ,$$

missä

1. **U** on $m\times n$-matriisi (siis samaa muotoa kuin **A**) ja pystyriveittäin ortonormaalinen eli $\mathbf{U}'\mathbf{U}=\mathbf{I}$,
2. **V** on ortogonaalinen $n\times n$-matriisi (eli $\mathbf{V}'\mathbf{V}=\mathbf{I}$ ja $\mathbf{VV}'=\mathbf{I}$),
3. **D** on ei-negatiivisten singulaariarvojen $d_1 \geq d_2 \geq ... \geq d_n \geq 0$ muodostama lävistäjämatriisi.

Todistus. Olkoon matriisin **A** aste r(**A**)=r ($r\leq n$). Tällöin $\mathbf{A}'\mathbf{A}$ on ei-negatiivisesti definiitti ja sillä on spektraalihajotelma

$$(1)\quad \mathbf{A}'\mathbf{A} = \mathbf{VD}^2\mathbf{V}' ,$$

missä $\mathbf{D}^2$ on $n\times n$-lävistäjämatriisi

$$\mathbf{D}^2 = \begin{bmatrix} \mathbf{D}_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

$$\mathbf{D}_1^2 = \mathrm{diag}(d_1^2, d_2^2, ..., d_r^2), \quad d_1^2 \geq d_2^2 \geq ... \geq d_r^2 > 0 .$$

Ortogonaalinen $n\times n$-matriisi **V** ositetaan muotoon

$$\mathbf{V} = [\, \mathbf{V}_1 \;\; \mathbf{V}_2 \,] ,$$

missä $n\times r$-matriisi $\mathbf{V}_1$ koostuu $\mathbf{A}'\mathbf{A}$:n positiivisia ominaisarvoja vastaavista ominaisvektoreista ja $n\times(n-r)$-matriisi $\mathbf{V}_2$ nolla-ominaisarvoja vastaavista ominaisvektoreista. Matriisi **V** on lisäksi ortogonaalinen eli

$$\mathbf{VV}' = \mathbf{V}'\mathbf{V} = \mathbf{I} .$$

Tällöin (1) voidaan kirjoittaa muodossa

$$\mathbf{A}'\mathbf{A} = \mathbf{V}_1\mathbf{D}_1^2\mathbf{V}_1' ,$$

mistä seuraa myös, että

$$(2)\quad \mathbf{D}_1^2 = \mathbf{V}_1'\mathbf{A}'\mathbf{A}\mathbf{V}_1 .$$

Olkoon $m\times r$-matriisi $\mathbf{U}_1$ määritelty siten, että

$$(3)\quad \mathbf{U}_1 = \mathbf{A}\mathbf{V}_1\mathbf{D}_1^{-1} .$$

Tällöin matriisin $\mathbf{U}_1$ sarakkeet ovat ortonormaalisia, sillä (2):n perusteella

$$\mathbf{U}_1'\mathbf{U}_1 = \mathbf{D}_1^{-1}\mathbf{V}_1'\mathbf{A}'\mathbf{A}\mathbf{V}_1\mathbf{D}_1^{-1} = \mathbf{D}_1^{-1}\mathbf{D}_1^2\mathbf{D}_1^{-1} = \mathbf{I} .$$

Tunnetusti, kun $m\geq n$, on mahdollista muodostaa sellainen $m\times(m-r)$-matriisi $\mathbf{U}_2$, että

$\mathbf{U}^* = [\ \mathbf{U}_1\ \ \mathbf{U}_2\ ]$

on ortogonaalinen. Kun nyt (3) kerrotaan oikealta matriisilla $\mathbf{D}_1\mathbf{V}_1'$, saadaan matriisille **A** esitys

$$\mathbf{A} = \mathbf{U}_1\mathbf{D}_1\mathbf{V}_1' = [\ \mathbf{U}_1\ \ \mathbf{U}_2\ ] \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1' \\ \mathbf{V}_2' \end{bmatrix}$$

eli

(4) $\mathbf{A} = \mathbf{U}^*\mathbf{D}^*\mathbf{V}'$,

missä $\mathbf{D}^*$ on $m \times n$-matriisi. Koska $\mathbf{D}^*$:n m-n viimeistä vaakariviä ovat nollia, **A** voidaan myös esittää muodossa

(5) $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}'$,

missä **U** koostuu $\mathbf{U}^*$:n n ensimmäisestä pystyrivistä. Näin esityksessä (5) **U**,**V** ja **D** täyttävät väitteen ehdot. Myös "lavennettu" esitys (4), jossa $\mathbf{U}^*$ on ortogonaalinen $m \times m$-matriisi, on hyödyksi joissain tarkasteluissa.

Hajotelma voidaan kirjoittaa myös summamuodossa matriisien **U** ja **V** pystyvektoreiden avulla seuraavasti:

$$\mathbf{A} = d_1\mathbf{u}^{(1)}\mathbf{v}^{(1)'} + d_2\mathbf{u}^{(2)}\mathbf{v}^{(2)'} + ... + d_n\mathbf{u}^{(n)}\mathbf{v}^{(n)'} .$$

Hajotelmasta $\mathbf{A}=\mathbf{U}\mathbf{D}\mathbf{V}'$ seuraa suoraan, että

$\mathbf{A}\mathbf{A}' = \mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{V}\mathbf{D}\mathbf{U}' = \mathbf{U}\mathbf{D}^2\mathbf{U}'$.

Siis $\mathbf{u}^{(1)}$, $\mathbf{u}^{(2)}$, ..., $\mathbf{u}^{(n)}$ ovat matriisin $\mathbf{A}\mathbf{A}'$ ominaisvektorit.

Edellä on oletettu, että matriisin **A** vaakarivien lukumäärä m on suurempi tai yhtäsuuri kuin pystyrivien lukumäärä n. Päinvastaisessa tapauksessa SVD on sama kuin mitä saadaan transponoimalla **A**:n SVD eli muotoa

$\mathbf{A}' = \mathbf{V}\mathbf{D}\mathbf{U}'$.

Singulaariarvohajotelmaa ei käytännössä kannata muodostaa $\mathbf{A}'\mathbf{A}$:n spektraalihajotelman kautta, vaan on sovellettavissa tehokkaampia laskentamenetelmiä, joissa esim. **A** ensin redusoidaan ns. bidiagonaalimuotoon ja tämä edelleen diagonalisoidaan iteratiivisesti. Survon matriisitulkissa ja matriisialiohjelmissa on käytetty *G.H.Golubin* ja *C.Reinschin* v. 1971 julkaisemaa algoritmia, joka on nopea sekä numeerisesti erittäin tarkka ja stabiili.

Tulokset ovat yleensä yhtä tarkkoja, mikä on koneessa liukulukujen esitystarkkuus (PC:ssä siis noin 15-16 merkitsevää numeroa). Ei siis synny juuri kasautuvia pyöristysvirheitä. Matriisin vajaa-asteisuudesta tai likimaisista lineaarisista riippuvuuksista ei ole mitään haittaa. Vaikeita ovat vain sellaiset patologiset tapaukset, joissa alkioiden suuruus vaihtelee kohtuuttomasti; esim.

yhtaikaa esiintyy matriisin alkioita, jotka ovat suuruusluokkaa 10^{20} ja 10^{-20}. Tällaisesta voi välttyä kaikissa järkevästi skaalatuissa tilastolllisissa sovelluksissa.

SVD:n ominaisuuksia

Matriisin singulaariarvohajotelma tulee vastaan useissa monimuuttujamenetelmissä. Erityisesti tietyt optimaalisuusominaisuudet ovat keskeisessä asemassa.

Lineaarisen kuvauksen yleinen rakenne
Tarkasteltaessa lineaarista kuvausta **y=Ax** havaitsemme ottamalla käyttöön **A**:n SVD:n, että **y=UDV'x**, mikä merkitsee sitä, että jokainen lineaarinen kuvaus voidaan nähdä yksinkertaisesti koordinaatiston kierron, akselien suuntaisten venytysten ja kutistusten sekä toisen koordinaatiston kierron yhdistelmänä. Tätä seikkaa hyödynnämme multinormaalijakauman perusominaisuuksia tarkasteltaessa.

Bilineaarimuodon x'Ay maksimointi
Seuraavasta tuloksesta on hyötyä mm. kanonisten korrelaatioiden ja korrespondenssianalyysin yhteydessä.

Funktion $f(\mathbf{x},\mathbf{y})=\mathbf{x}'\mathbf{A}\mathbf{y}$, missä **A** on $m\times n$-matriisi ja **x** ja **y** yksikkövektorin mittaisia vektoreita, maksimiarvo on **A**:n SVD:n $\mathbf{A}=\mathbf{U}\mathbf{D}\mathbf{V}'$ suurin singulaariarvo d_1 ja maksimi saavutetaan kun $\mathbf{x}=\mathbf{u}^{(1)}$ ja $\mathbf{y}=\mathbf{v}^{(1)}$.

Todistus. Käytetään "lavennettua" esitystä (4) $\mathbf{A}=\mathbf{U}^*\mathbf{D}^*\mathbf{V}'$, jolloin jokainen **x**-vektori voidaan lausua muodossa

$$\mathbf{x} = a_1\mathbf{u}^{(1)} + a_2\mathbf{u}^{(2)} + ... + a_m\mathbf{u}^{(m)} = \mathbf{U}^*\mathbf{a}\ ,\ \text{missä}\ \mathbf{a}'\mathbf{a} = 1\ ,$$

koska $\mathbf{U}^*$-vektorit muodostavat m-ulotteisen avaruuden ortonormeeratun kannan. Jokainen **y**-vektori voidaan esittää vastaavasti muodossa

$$\mathbf{y} = b_1\mathbf{v}^{(1)} + b_2\mathbf{v}^{(2)} + ... + b_n\mathbf{v}^{(n)} = \mathbf{V}\mathbf{b}\ ,\ \text{missä}\ \mathbf{b}'\mathbf{b} = 1\ .$$

Tällöin on

$$\begin{aligned} f(\mathbf{x},\mathbf{y}) &= \mathbf{a}'\mathbf{U}^{*\prime}\mathbf{U}^*\mathbf{D}^*\mathbf{V}'\mathbf{V}\mathbf{b} = \mathbf{a}'\mathbf{D}^*\mathbf{b} \\ &= a_1b_1d_1 + a_2b_2d_2 + ... + a_nb_nd_n\ , \end{aligned}$$

mikä maksimoituu ehdoilla $\mathbf{a}'\mathbf{a}=\mathbf{b}'\mathbf{b}=1$, kun

$a_1=1,\ a_2=a_3=...=a_m=0$
$b_1=1,\ b_2=b_3=...=b_n=0$

ja maksimiarvoksi saadaan suurin singulaariarvo d_1. Maksimin antavat **x** ja **y** ovat $\mathbf{u}^{(1)}$ ja $\mathbf{v}^{(1)}$, kuten väitettiin.

Tämä osoitetaan tarkasti soveltamalla Schwarzin epäyhtälöä

$(\mathbf{s}'\mathbf{t})^2 \le (\mathbf{s}'\mathbf{s})(\mathbf{t}'\mathbf{t})$,

kun $\mathbf{s}=\mathbf{a}$ ja $\mathbf{t}=\mathbf{D}^*\mathbf{b}$, jolloin

$$(a_1b_1d_1 + ... + a_nb_nd_n)^2 = (\mathbf{a}'\mathbf{D}^*\mathbf{b})^2 \le (\mathbf{a}'\mathbf{a})(\mathbf{b}'\mathbf{D}^{*\prime}\mathbf{D}^*\mathbf{b}) = b_1^2d_1^2 + ... + b_n^2d_n^2 ,$$

mikä painotetuksi keskiarvoksi tulkittuna on $\le d_1^2$.

Vastaavasti nähdään suoraan, että lisäehdoilla

$\mathbf{x}'\mathbf{u}^{(i)} = 0$ ja/tai $\mathbf{y}'\mathbf{v}^{(i)} = 0$, $i = 1,2,...,k\text{-}1$

eli

$a_i = 0$ ja/tai $b_i = 0$, $i = 1,2,...,k\text{-}1$

$f(\mathbf{x},\mathbf{y})=\mathbf{x}'\mathbf{A}\mathbf{y}$ saavuttaa maksiminsa, joka on d_k, kun $\mathbf{x}=\mathbf{u}^{(k)}$ ja $\mathbf{y}=\mathbf{v}^{(k)}$ kaikilla arvoilla $k=2,3,...,n$.

Äskeinen tulos on analoginen yleisesti tunnetun yleisen neliömuodon maksimointia koskevan kanssa: $\mathbf{x}'\mathbf{A}\mathbf{x}$, missä $\mathbf{A}$ on symmetrinen neliömatriisi ja $\mathbf{x}$ on yksikkövektorin mittainen vektori, saavuttaa maksiminsa, joka on $\mathbf{A}$:n suurin ominaisarvo, kun $\mathbf{x}$ on tätä ominaisarvoa vastaava ominaisvektori. Todistus on edellisen kaltainen.

Matriisin approksimointi alempiasteisella matriisilla

Eräissä monimuuttujamenetelmissä joudutaan tilanteeseen, jossa esim. kovarianssimatriisille Σ joudutaan etsimään tietynrakenteinen, alempiasteinen approksimaatio. Näin tapahtuu mm. pääkomponentti- ja faktorianalyysissa.

Tutkittaessa, miten hyvin esim. matriisi $\mathbf{B}$ vastaa matriisia $\mathbf{A}$, käytetään erotusmatriisin $\mathbf{A}\text{-}\mathbf{B}$ koon mittana sen alkioiden neliösummaa. Tätä neliösummaa merkitään $\|\mathbf{A}\text{-}\mathbf{B}\|^2$.

Yleisestikin matriisin $\mathbf{A}$ kokoa mitataan tällä ns. Frobeniuksen normilla

$$\|\mathbf{A}\|^2 = \mathrm{tr}(\mathbf{A}\mathbf{A}') = \mathrm{tr}(\mathbf{A}'\mathbf{A}) = \sum_i \sum_j a_{ij}^2 .$$

Vektoreilla $\|\mathbf{x}\|$ tarkoittaa yksinkertaisesti vektorin pituutta.

$m\times n$-matriisin $\mathbf{A}$ singulaariarvohajotelmasta $\mathbf{A}=\mathbf{U}\mathbf{D}\mathbf{V}'$ seuraa, että yleisesti

$$\|\mathbf{A}\|^2 = \mathrm{tr}(\mathbf{A}\mathbf{A}') = \mathrm{tr}(\mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{V}\mathbf{D}\mathbf{U}') = \mathrm{tr}(\mathbf{D}^2\mathbf{U}'\mathbf{U}) = d_1^2 + ... + d_n^2 .$$

Jos nyt haluamme approksimoida matriisia **A** toisella $m \times n$-matriisilla **B**, joka on (alempaa) astetta r(**B**)=$p<n$, tuntuu luonnolliselta, kun ottaa huomioon **A**:n SVD:n muodossa

$$\mathbf{A} = d_1\mathbf{u}^{(1)}\mathbf{v}^{(1)'} + d_2\mathbf{u}^{(2)}\mathbf{v}^{(2)'} + ... + d_n\mathbf{u}^{(n)}\mathbf{v}^{(n)'} ,$$

että

$$(6)\quad \mathbf{B} = d_1\mathbf{u}^{(1)}\mathbf{v}^{(1)'} + d_2\mathbf{u}^{(2)}\mathbf{v}^{(2)'} + ... + d_p\mathbf{u}^{(p)}\mathbf{v}^{(p)'}$$

on hyvä ehdokas tällaiseksi matriisiksi.

Yritämmekin näyttää toteen, että Frobeniuksen normin mielessä $\|\mathbf{A}\text{-}\mathbf{B}\|^2$ saavuttaa minimiarvon, joka on

$$d_{p+1}^2 + ... + d_n^2 ,$$

kun **B** on edellä mainittu. Itse asiassa vastaavaan tulokseen päädytään monilla muillakin matriisinormeilla. Todistuksemme seuraa linjoja, jotka *Seppo Hassi* on esittänyt.

Toteamme aluksi, että jos **S** on ortogonaalinen $n \times n$-matriisi, matriisin **AS** normi on sama kuin **A**:n, sillä

$$\|\mathbf{AS}\|^2 = \mathrm{tr}(\mathbf{ASS'A'}) = \mathrm{tr}(\mathbf{AA'}) = \|\mathbf{A}\|^2 .$$

Osoitamme nyt ensin, että jos **S** ositetaan muotoon, $\mathbf{S} = [\ \mathbf{S}_1\ \mathbf{S}_2\ ]$, missä $\mathbf{S}_1$ on $n \times p$-matriisi ja siis pystyriveittäin ortogonaalinen eli $\mathbf{S}_1'\mathbf{S}_1=\mathbf{I}$, niin

$$(7)\quad \|\mathbf{AS}_1\|^2 \le d_1^2 + ... + d_p^2 .$$

Kirjoitamme tarkastelun kohteena olevan normin neliön muotoon

$$\|\mathbf{AS}_1\|^2 = \mathrm{tr}(\mathbf{S}_1'\mathbf{A'AS}_1) = \mathrm{tr}(\mathbf{S}_1'\mathbf{VD}^2\mathbf{V'S}_1) = \mathrm{tr}(\mathbf{C'D}^2\mathbf{C}) ,$$

missä **C** on $n \times p$-matriisi

$$\mathbf{C} = \mathbf{V'S}_1 .$$

Myös matriisi **C** on tällöin pystyriveittäin ortonormaalinen eli

$$\mathbf{C'C} = \mathbf{I} .$$

Merkitään

$$\mathbf{C'} = [\ \mathbf{c}^{(1)}\ \mathbf{c}^{(2)}\ ...\ \mathbf{c}^{(n)}\ ] ,$$

jolloin

$$(8)\quad \mathbf{c}^{(i)'}\mathbf{c}^{(i)} = g_i \le 1 ,\ i = 1, ..., n ,\ \ \mathrm{tr}(\mathbf{CC'}) = g_1 + ... + g_n = p .$$

Maksimointitehtävä palautuu tällöin lausekkeen

$$\mathrm{tr}(\mathbf{C'D}^2\mathbf{C}) = \mathrm{tr}(\mathbf{CC'D}^2) = g_1d_1^2 + ... + g_nd_n^2$$

maksimoinniksi ehdoilla (8). Tästä on helppo päätellä, että ko. maksimi saavutetaan, kun

$$g_1 = ... = g_p = 1 ,\ g_{p+1} = ... = g_n = 0$$

ja maksimiarvo on $d_1^2 + ... + d_p^2$.

Olkoon nyt $m \times n$ matriisin **B** (r(**B**)=p) SVD

$$\mathbf{B} = \mathbf{U}_B\mathbf{D}_B\mathbf{W}'$$

ja tarkastellaan **W**:n ositusta $\mathbf{W} = [\ \mathbf{W}_1\ \mathbf{W}_2\ ]$, missä $\mathbf{W}_1$ on $n \times p$-matriisi. Nyt voimme päätellä, että

$$\|\mathbf{A}\text{-}\mathbf{B}\|^2 = \|(\mathbf{A}\text{-}\mathbf{B})\mathbf{W}\|^2 = \|(\mathbf{A}\text{-}\mathbf{B})\mathbf{W}_1\|^2 + \|(\mathbf{A}\text{-}\mathbf{B})\mathbf{W}_2\|^2 \geq \|(\mathbf{A}\text{-}\mathbf{B})\mathbf{W}_2\|^2 ,$$

mutta

$$\mathbf{B}\mathbf{W}_2 = \mathbf{0} ,$$

koska **B**:n viimeiset n-p singulaariarvoa ovat nollia ja

$$\|\mathbf{A}\mathbf{W}_2\|^2 = \|\mathbf{A}\mathbf{W}\|^2 - \|\mathbf{A}\mathbf{W}_1\|^2 .$$

Koska

$$\|\mathbf{A}\mathbf{W}\|^2 = d_1^2 + ... + d_n^2 ,$$

ja (7):n mukaan

$$\|\mathbf{A}\mathbf{W}_1\|^2 \leq d_1^2 + ... + d_p^2 ,$$

on siis

$$\|\mathbf{A}\text{-}\mathbf{B}\|^2 \geq \|\mathbf{A}\mathbf{W}_2\|^2 \geq d_{p+1}^2 + ... + d_n^2$$

ja tämä alaraja saavutetaan, kun **B**:llä on esitys (6).

SVD:n "heuristinen" perustelu

Jos $m \times n$- (m>n) matriisi **A** on täysiasteinen, singulaarihajotelma saadaan johdettua hyvin suoraan seuraavasti:

A'A:lla on tällöin spektraalihajotelma

$$\mathbf{A}'\mathbf{A}=\mathbf{V}\mathbf{D}^2\mathbf{V}'$$

missä **V** on ortogonaalinen $n \times n$-matriisi ja $\mathbf{D}^2$ positiivisten ominaisarvojen muodostama lävistäjämatriisi. Määritellään

$$\mathbf{U}=\mathbf{A}\mathbf{V}\mathbf{D}^{-1} .$$

Tällöin

$$\mathbf{U}'\mathbf{U}=\mathbf{D}^{-1}\mathbf{V}'\mathbf{A}'\mathbf{A}\mathbf{V}\mathbf{D}^{-1}=\mathbf{D}^{-1}\mathbf{V}'\mathbf{V}\mathbf{D}^2\mathbf{V}'\mathbf{V}\mathbf{D}^{-1}=\mathbf{I}$$

eli **U** on pystyriveittäin ortogonaalinen $n \times m$-matriisi ja

$$\mathbf{A}=\mathbf{U}\mathbf{D}\mathbf{V}'.$$

Kirjallisuutta

Ahmavaara, Y. (1954). Transformation analysis of factorial data. *Ann.Acad.Sci.Fenn.*, **B88,2**.

Ahmavaara, Y. & Vahervuo, T. (1958). *Johdatus Faktorianalyysiin*. WSOY.

Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*, Wiley.

Andrews, D.F. (1972). Plots of high-dimensional data. *Biometrics*, **28**, 125-36.

Box, G.E.P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, **36**, 317-346.

Chatfield, C. & Collins, A.J. (1980). *Introduction to Multivariate Analysis*. Chapman & Hall.

Chernoff, H. (1973). Using faces to represent points in k-dimensional space graphically. *JASA*. **68**, 361-368.

Cox, T.F.C. & Cox, M.A.A. (1994). *Multidimensional Scaling*. Chapman & Hall.

Giri, N.C. (1977). *Multivariate Statistical Inference*. Academic Press.

Harman, H.H. (1967). *Modern Factor Analysis*, Second Edition. The University of Chicago Press.

Jennrich, R.I. & Sampson, P.F. (1966). Rotation for simple loadings. *Psych.*, **31**, 313-323.

Jöreskog, K.G. (1963). *Statistical Estimation in Factor Analysis*. Almqvist and Wiksell.

Kaiser, H.F. (1958). The Varimax Criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, 187-200.

Korhonen, P. (1979). *A Stepwise Procedure for Multivariate Grouping*. Computing Centre, University of Helsinki.

Lawley, D.N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proc.Roy.Soc.Edin*, **60**, 64-82.

Lebart, L., Morineau, A. & Warwick, K.M. (1984). *Multivariate Descriptive Statistical Analysis*. Wiley.

Ledermann, W. (1938). Shortened method of estimation of mental factors by regression. *Nature*, **141**, 650.

Mustonen, S. (1966). *Symmetrinen transformaatioanalyysi*. Alkoholipoliittisen tutkimuslaitoksen tutkimusseloste N:o 24.

Mustonen, S. (1992). *Survo, An Integrated Environment for Statistical Computing and Related Areas*. Survo Systems Ltd.

Nelder, J.A. & Mead, R. (1963). A simplex method for function minimzation. *The Computer Journal*, **7**, 308-313.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, Wiley.

Seber, G.A.F. (1985). *Multivariate Observations*, Wiley.

Torgerson, W.S. (1952). Multidimensional scaling: I-theory and method. *Psychometrika*, **17**, 401-419.