Monimuuttujamenetelmät / Kimmo Vehkalahti

Tämä luentomoniste on alunperin tarkoitettu oheismateriaaliksi n. 15 tunnin mittaiselle sovelluspainotteiselle **monimuuttujamenetelmien** kurssille. Olen pitänyt useita sellaisia kursseja joko useamman viikon kestävänä periodiopetuksena yliopistolla tai parin päivän jaksona erilaisissa tutkimuslaitoksissa.

Itseopiskeluun materiaalista lienee hyötyä lähinnä niille, jotka tuntevat aihepiiriä jo entuudestaan tai haluavat kerrata aiemmin oppimaansa. Kurssin aikana tulee tyypillisesti esille arviolta kaksinkertainen määrä tietoa "rivien välistä", joten tämä esitys ei ole eikä yritäkään olla mitenkään tyhjentävä. Lisäksi kurssilaiset voivat tuoda omia kysymyksiään ja eri alojen sovellustilanteita käsiteltäviksi ja keskusteltaviksi. Näin jokainen kurssi on aina hieman erilainen. Tähän monisteeseen olenkin pyrkinyt tiivistämään vain keskeisimpiä asioita.

Kurssin ydinkohtia:

- moniulotteisten ilmiöiden ja etäisyyksien mittaaminen
- keskeiset tilastolliset monimuuttujamenetelmät
- menetelmien yleiset oletukset ja rajoitukset
- menetelmille ominaiset graafiset tarkastelut

Käsiteltäviä menetelmiä:

- faktorianalyysi
- pääkomponenttianalyysi
- erotteluanalyysi
- ryhmittelymenetelmät
- moniulotteinen skaalaus
- korrespondenssianalyysi

Asioita lähestytään soveltajan näkökulmasta. Taustalla olevaa tilastotieteen teoriaa esitellään tarpeen mukaan. Tärkeintä on oppia valitsemaan tilanteeseen sopivia menetelmiä ja käyttämään niitä tarkoituksenmukaisesti sekä tulkitsemaan ohjelmien antamia tulostuksia oikealla tavalla.

Tilastotieteen perusasiat ja -käsitteet on hyvä hallita etukäteen. Myöskään ohjelmien käyttöä ei kurssilla opeteta, joten käytännön hyötyä ajatellen jonkin tarkoitukseen sopivan ohjelmiston (esimerkiksi Survo, SAS, SPSS, S-Plus, SYSTAT) hallinta on suotavaa. Kurssin aikana asioita havainnollistetaan Survon Windows-version **SURVO MM** avulla (ks. *www.survo.fi)*.

Kirjallisuutta:

Flury, B. (1997). A First Course in Multivariate Statistics.
Springer-Verlag, New York.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. (1998).
Multivariate Data Analysis. 5th ed., Prentice Hall.

Krzanowski, W. J. (2000). Principles of Multivariate Analysis
(revised edition). Oxford University Press.

Mustonen, S. (1995). Tilastolliset monimuuttujamenetelmät.
Survo Systems, Helsinki.

Ranta, E., Rita, H., & Kouki, J. (1991). Biometria - tilastotiedettä ekologeille
(3. painos). Yliopistopaino, Helsinki.

Sivumennen sanoen niin tekstin kirjoittamisen, aineistojen analyysit, laskutoimitukset ja kuvien piirtämiset kuin ulkoasun viimeistelynkin olen tehnyt Survolla. Sen ansiosta moniste syntyikin nopeasti, vain parissa päivässä ennen Metsäntutkimuslaitoksella keväällä 2002 pitämääni kurssia.

Kaikki palaute on tervetullutta!

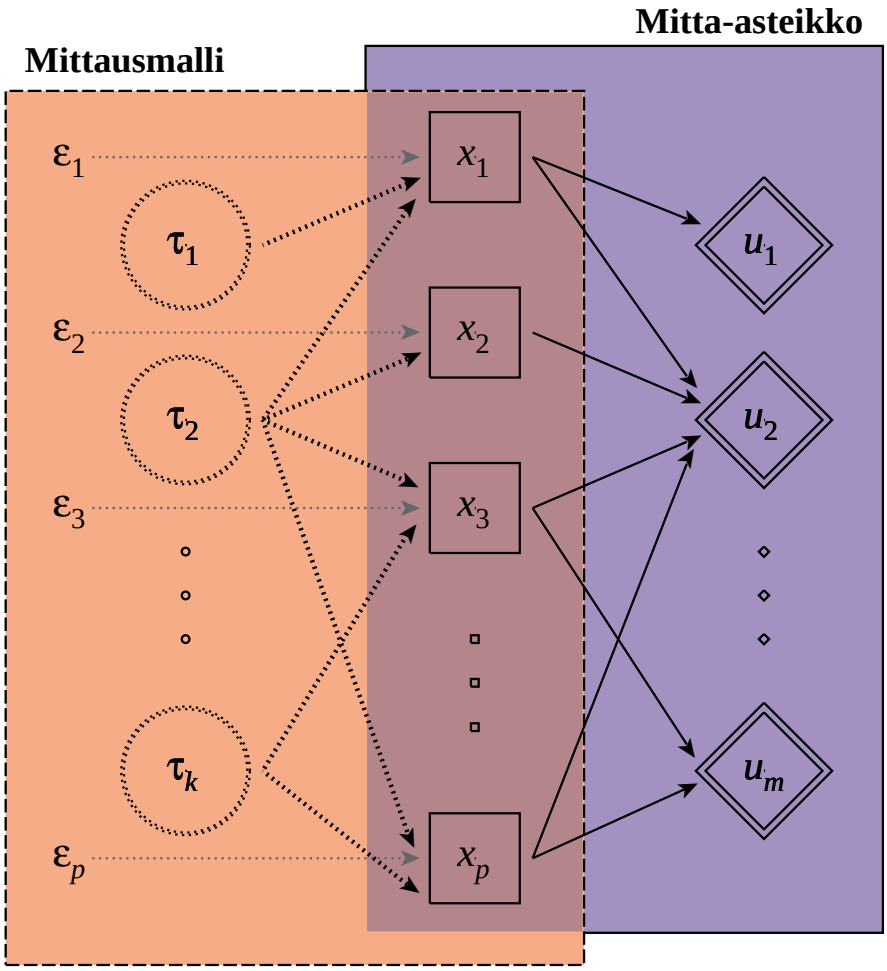
Kimmo Vehkalahti

http://www.helsinki.fi/people/Kimmo.Vehkalahti/ *Kimmo.Vehkalahti@helsinki.fi*

Faktorianalyysi

Faktorianalyysin *(factor analysis, FA)* perustana on tilastollinen malli, jossa ajatellaan havaittujen muuttujien riippuvuusrakenteen ilmentävän varsinaisen mielenkiinnon kohteena olevia piilomuuttujia joita ei voi suoraan havaita. Näitä ns. *latentteja* muuttujia kutsutaan faktoreiksi.

Riippuvuusrakenne on mielekästä hahmotella oheisen kuvan mukaisena mittausmallina *(measurement model)*. Tavoitteena on luoda uusia muuttujia asteikkoina *(measurement scales)*, jotka kuvaavat teoreettisia faktoreita mahdollisimman hyvin ja sisältävät samalla mahdollisimman vähän mittausvirheestä *(measurement error)* johtuvaa vaihtelua.



Mittauksen laadun arviointi on ensiarvoisen tärkeää kaikessa tieteellisessä tutkimuksessa. Ensisijalla ovat erilaiset validiteettitarkastelut, mutta myös mittarin tekninen tarkkuus on tärkeää. Sitä arvioidaan reliabiliteetin avulla. Asialliset reliabiliteettitarkastelut edellyttävät mittausmallin estimointia ja riittävän yleistä mitta-asteikkoa. Mittausmallin tärkeä erikoistapaus on faktorimalli.

Faktorianalyysi on eräs vanhimmista tilastollisista menetelmistä. Se sai alkunsa käyttäytymistieteiden puolella, mutta muotoutui sittemmin täysin yleiseksi tilastolliseksi monimuuttujamenetelmäksi, jota voidaan soveltaa mitä moninaisimmilla aloilla.

Tutkitaan esimerkkinä ihmisen fyysistä suorituskykyä. Oletetaan että se koostuu ainakin kolmesta komponentista: nopeus, voima ja kestävyys. Käytetään tässä mittarina kymmenottelua ja aineistona vuoden 1973 maailman 48 parhaan urheilijan saavuttamia lajipisteitä.

```
FILE STATUS KYMMEN
    Parhaat kymmenottelijat vuonna 1973
FIELDS: (active)
   1 SA-   8 Nimi     Ottelijan nimi
   2 NA-   2 Pisteet  Yhteispisteet
   3 NA-   2 100m     100 m juoksu (pisteet)
   4 NA-   2 Pituush  Pituushyppy
   5 NA-   2 Kuula    Kuulantyöntö
   6 NA-   2 Korkeus  Korkeushyppy
   7 NA-   2 400m     400 m juoksu
   8 NA-   2 Aidat    110 m aidat
   9 NA-   2 Kiekko   Kiekonheitto
  10 NA-   2 Seiväs   Seiväshyppy
  11 NA-   2 Keihäs   Keihäänheitto
  12 NA-   2 1500m    1500 m juoksu
  13 NA-   2 Pituus   Pituus (cm)
  14 NA-   2 Paino    Paino (kg)
END
Survo data file KYMMEN: record=128 bytes, M1=30 L=64  M=14 N=48
```

Valitaan analyysiin lajimuuttujat ja lasketaan niiden keskiarvot, hajonnat ja korrelaatiot.

```
MASK=--AAAAAAAAAA--
CORR KYMMEN
```

Tarkistetaan keskiarvot ja hajonnat sekä täydellisten havaintojen lukumäärät muuttujittain. Mikäli havaintoarvoja puuttuu, koko havainto jää käsittelyjen ulkopuolelle (ns. *listwise deletion* -periaate). Siis vain täydelliset havainnot kelpuutetaan. Toinen, parittaisiin tietoihin perustuva korrelaatioiden laskentamenetelmä voi antaa hyvin harhaisia tuloksia, joten sen käyttö ei ole suositeltavaa. Tiedon puuttumisen syyt pitää selvittää. Puuttuminen ei ole välttämättä satunnaista. Se voi olla systemaattista jonkin muun asian suhteen. Tiedon puuttuminen voi myös johtua siitä ettei kaikkia tietoja ole ollut tarkoituskaan mitata kaikilta havaintoyksiköiltä.

Aidosti puuttuvia tietoja ei mikään korvaa, mutta eri asteisia paikkauksia voi olla mahdollista tehdä ja saada jonkin verran vältettyä havaintojen täydellistä menettämistä. Usein käytännössä sovellettu keskiarvolla korvaaminen on melko alkeellista, ja sitä pitäisi välttää jos parempia menetelmiä on käytettävissä. Parhaat keinot perustuvat regressioestimointiin, ja ovat hyvinkin tehokkaita.

Tässä aineistossa havainnot ovat täydellisiä, joten paikkausta ei tarvita.

```
/LOADMSN KYMMEN
Means, standard deviations and number of observations in KYMMEN:
            mean  stddev   N
100m     828.188  59.303  48 100 m juoksu (pisteet)
Pituush  840.188  50.729  48 Pituushyppy
Kuula    740.771  61.828  48 Kuulantyöntö
Korkeus  805.854  64.805  48 Korkeushyppy
400m     813.500  49.802  48 400 m juoksu
Aidat    852.875  54.205  48 110 m aidat
Kiekko   747.458  62.282  48 Kiekonheitto
Seiväs   900.271  63.043  48 Seiväshyppy
Keihäs   760.021  63.937  48 Keihäänheitto
1500m    554.625  76.672  48 1500 m juoksu
```

Vilkaistaan korrelaatiomatriisia:

```
/LOADCORR

Limits: P=0.001 0.46   P=0.01 0.364   P=0.05 0.283
R(KYMMEN)
          100m Pituus  Kuula Korkeu   400m  Aidat Kiekko Seiväs Keihäs  1500m
100m     1.000  0.172 -0.028 -0.412  0.456  0.316  0.014  0.055 -0.221 -0.292
Pituush  0.172  1.000 -0.034 -0.003  0.133  0.298  0.021  0.061  0.154 -0.207
Kuula   -0.028 -0.034  1.000  0.163 -0.304  0.086  0.727 -0.204  0.023 -0.446
Korkeus -0.412 -0.003  0.163  1.000 -0.339 -0.039  0.217 -0.118  0.150 -0.146
400m     0.456  0.133 -0.304 -0.339  1.000  0.176 -0.345  0.007 -0.105  0.302
Aidat    0.316  0.298  0.086 -0.039  0.176  1.000  0.048 -0.073 -0.148 -0.225
Kiekko   0.014  0.021  0.727  0.217 -0.345  0.048  1.000 -0.182  0.136 -0.574
Seiväs   0.055  0.061 -0.204 -0.118  0.007 -0.073 -0.182  1.000 -0.129  0.012
Keihäs  -0.221  0.154  0.023  0.150 -0.105 -0.148  0.136 -0.129  1.000 -0.065
1500m   -0.292 -0.207 -0.446 -0.146  0.302 -0.225 -0.574  0.012 -0.065  1.000
```

Korrelaatioistakin näkyy jo yhtä ja toista, esim. suurin korrelaatio (0.727) on kiekonheiton ja kuulantyönnön välillä. Korrelaatiomatriisi on kuitenkin vain lähtökohta useille monimuuttujamenetelmille; siitä ei pidä tehdä liian pitkälle meneviä päätelmiä. Varsinkaan ei kannata tuijottaa yksittäisten korrelaatioiden tilastollisiin merkitsevyyksiin (jotka edellä on kuitenkin automaattisesti korostettu yleissilmäilyn helpottamiseksi). Merkitsevyyksiä ei pidä ylipäätään ottaa liian vakavasti. On muistettava että tilastollinen merkitsevyys riippuu otoskoosta: suurilla aineistoilla kaikki on tilastollisesti merkitsevää *(significant)* vaikkei välttämättä käytännössä lainkaan merkittävää *(notable, remarkable)*.

Kun muuttujia on enemmän, lukujen silmäilykin käy äkkiä hankalaksi. Vaikka keskiarvot, hajonnat ja korrelaatiot (ns. tyhjentävät otossuureet) tiivistävätkin jo aineiston tietoa melkoisesti, ei se kuitenkaan riitä vielä mihinkään. Tarvitaan menetelmiä joilla informaatiota survotaan tästä huomattavasti tiiviimmäksi paketiksi.

Faktorianalyysi on usein jo tutkimuksen alkuvaiheessa sovellettu menetelmä, jolla saadaan tarkasteltavien muuttujien määrä realistisemmaksi. Samalla saadaan erotettua todellinen vaihtelu satunnaisesta mittausvirhevaihtelusta, mikä antaa mahdollisuuden arvioida mm. uusien muuttujien reliabiliteettia ja mittauksen keskivirhettä. Näin jatkoanalyysit voidaan tehdä muuttujilla joista on puhdistettu mittausvirheiden vaikutus.

Oikean faktoriluvun määrääminen on olennaisen tärkeää. Sitä ei saa antaa ohjelman (korrelaatiomatriisin ominaisarvojen perusteella) "keksiä" vaan sen on oltava tutkijan vastuulla. Tutkijanhan parhaiten luulisi tietävän, minkälaista ja miten moniulotteista ilmiötä on mallintamassa!

Ennakkokäsityksen mukaan tutkittava ilmiö on (ainakin) kolmiulotteinen, joten faktoroidaan korrelaatiomatriisi sen mukaisesti käyttäen faktorilukuna kolmea. Sovelletaan suurimman uskottavuuden *(maximum likelihood)* faktorointimenetelmää. Se on käytännössä suositeltavin. Muita asiallisia vaihtoehtoja ovat lähinnä pääakselimenetelmä *(principal axes)* tai yleistetty pienimmän neliösumman menetelmä *(generalized least squares)*. Historiallisista syistä monissa ohjelmissa esiintyy vaihtoehtona (jopa oletuksena) tässä yhteydessä pääkomponenttianalyysi *(principal components)*, mutta se ei ole sama asia kuin faktorianalyysi, joten sitä pitää osata tietoisesti välttää, jos haluaa tehdä kunnollista faktorianalyysia. Muut mahdollisesti tarjolla olevat vaihtoehdot kuten esim. alfa-faktorointi yms. ovat jäänteitä historiasta.

Tehdään siis faktorointi kolmella faktorilla edellä olevasta korrelaatiomatriisista.

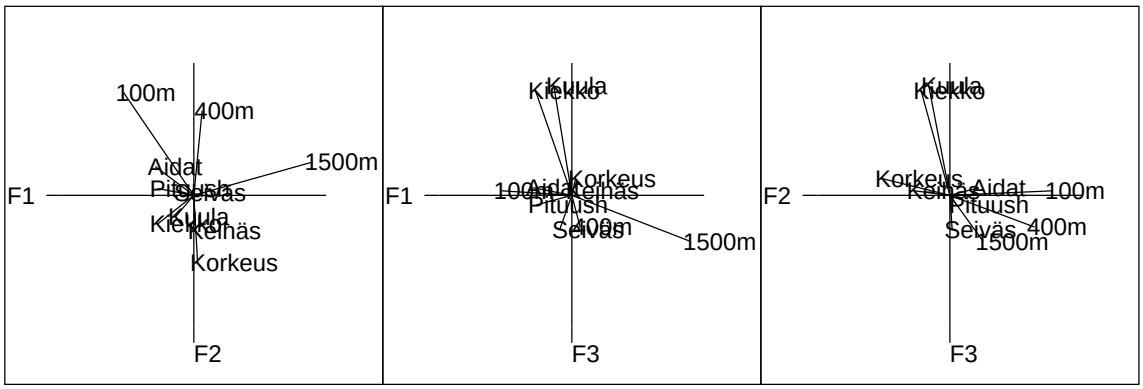
```
FACTA CORR.M,3,CUR+1
Factor analysis: Maximum Likelihood (ML) solution
Factor matrix
             F1     F2     F3    h^2
100m     -0.298  0.875  0.176  0.886
Pituush  -0.206  0.163 -0.112  0.082
Kuula    -0.456 -0.313  0.654  0.733
Korkeus  -0.144 -0.501 -0.061  0.275
400m      0.300  0.617  0.035  0.471
Aidat    -0.227  0.283  0.058  0.135
Kiekko   -0.582 -0.301  0.562  0.745
Seiväs    0.016  0.115 -0.245  0.073
Keihäs   -0.064 -0.254 -0.058  0.072
1500m     0.997  0.004  0.014  0.995
```

Tulkinnan selkiyttämiseksi suoritetaan saadulle faktorimatriisille ortogonaalinen Varimax-rotaatio ja otetaan lopputulos esille siten että tulkinnan perusteet ovat selvästi näkyvissä. Tavoitteena on ns. yksinkertainen rakenne *(simple structure)*. Asiaa voisi lähestyä tarkemminkin graafisen rotaation avulla.

```
ROTATE FACT.M,3
/LOADFACT KYMMEN
             F3     F2     F1  Sumsqr
Kuula     0.831 -0.160 -0.132  0.733 Kuulantyöntö
Kiekko    0.791 -0.213 -0.274  0.745 Kiekonheitto
100m      0.038  0.778 -0.529  0.886 100 m juoksu (pisteet)
400m     -0.235  0.642  0.064  0.471 400 m juoksu
Korkeus   0.126 -0.509  0.030  0.275 Korkeushyppy
1500m    -0.345  0.255  0.901  0.995 1500 m juoksu
Pituush  -0.069  0.056 -0.272  0.082 Pituushyppy
Aidat     0.060  0.220 -0.288  0.135 110 m aidat
Seiväs   -0.255  0.020 -0.089  0.073 Seiväshyppy
Keihäs    0.036 -0.266  0.016  0.072 Keihäänheitto
Sumsqr    1.582  1.534  1.353  4.469
```

Faktorit näyttäisivät löytyvän ennakkokäsityksen mukaisesti järjestyksessä F1: kestävyys, F2: nopeus, F3: voima (taulukon järjestys faktoreiden voimakkuuksien mukaan). Siis esim. kuulantyöntö ja kiekonheitto latautuvat kolmannelle faktorille, jossa suurin negatiivinen korrelaatio on 1500 metrin juoksulla. Tulkinta on varsin selvä: voimalajeissa menestyvät ovat isokokoisina vaikeuksissa kestävyysjuoksussa.

Havainnollistetaan rotatoitua faktoriratkaisua graafisesti piirtämällä faktoriavaruuden dimensiot pareittain vastakkain. Muuttujat esiintyvät faktoriavaruudessa vektoreina, jotka kuvaavat vastaavia faktorilatauksia eli korrelaatioita faktorien ja muuttujien välillä.



Palataan havaintomatriisin tasolle estimoimalla havaintokohtaiset arvot faktoreittain eli ns. faktoripistemäärät *(factor scores)*. Tämä tapahtuu regressiomenetelmällä, sillä faktorianalyysin perusyhtälöä ei voi ratkaista yksikäsitteisesti faktoreiden suhteen.

Lasketaan aluksi tarvittava painokerroinmatriisi. Kertoimet ovat pieniä, koska muuttujien saamat arvot ovat suuria. Vakiotermillä *(Constant)* aikaansaadaan keskistys, eli faktoripistemäärien keskiarvot tulevat olemaan nollia.

```
/FCOEFF

Factor_score_coefficients
///            %1       %2       %3
Constant  -7.9920 -18.8213 -14.4382
100m      -0.0033   0.0141   0.0021
Pituush   -0.0003   0.0001  -0.0006
Kuula      0.0030   0.0016   0.0089
Korkeus    0.0003  -0.0011   0.0000
400m      -0.0007   0.0023  -0.0004
Aidat     -0.0002   0.0007   0.0001
Kiekko     0.0027   0.0012   0.0080
Seiväs    -0.0003  -0.0002  -0.0009
Keihäs     0.0001  -0.0005  -0.0001
1500m      0.0134   0.0071   0.0030
```

Lasketaan uudet muuttujat vanhojen lineaarikombinaatioina.

```
LINCO KYMMEN,FCOEFF.M(kestäv,nopeus,voima)
```

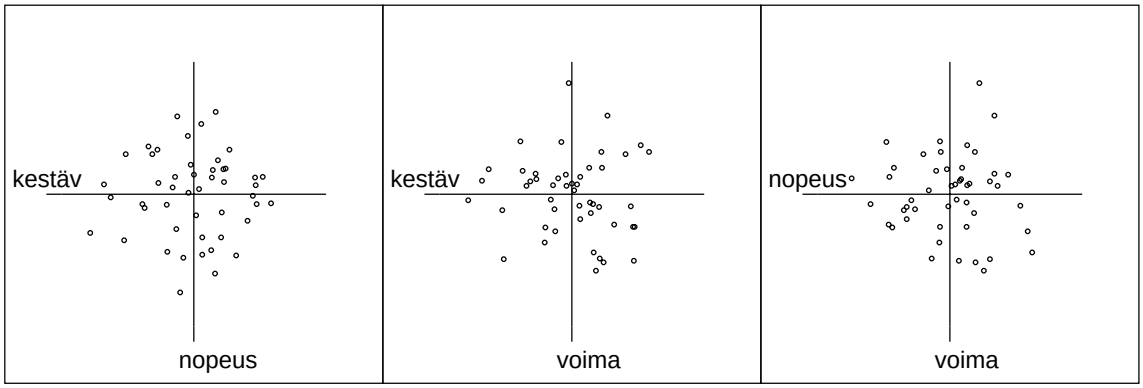
Lineaarikombinaatiot ovat siis painotettuja summia, joiden painot määräytyvät faktorianalyysin perusteella. Tällaiset muuttujat ovat monesta syystä suositeltavampia kuin ns. summamuuttujat, joissa muuttujille annetaan painoja 0 ja 1 osittain mielivaltaisesti.

Nyt kullekin urheilijalle on saatu kymmenen lajipisteen sijasta kolme arvoa, jotka kuvaavat fyysisen suorituskyvyn eri dimensioita, kestävyyttä, nopeutta ja voimaa. Kukin muuttujista on asteikko dimension ääripäästä toiseen. Sinänsä lukuarvot ovat anonyymejä, vaihdellen nollan molemmin puolin.

Parhaat urheilijat eri dimensioilla ovat Ghesquir (kestävyys), Bennett (nopeus) ja Zigert (voima).

```
FILE LOAD KYMMEN CUR+2 / VARS=Nimi,kestäv,nopeus,voima

 Nimi          kestäv      nopeus       voima
 Skowrone      -0.478       0.168      -0.106
 Hedmark       -1.157      -0.213       1.215
 Le_Roy        -1.876      -0.055       0.580
 Zeilbaue      -0.104       0.049       0.198
 Zigert        -0.067       0.682      2.532
 Bennett        0.501      1.869       -1.314
 Blinjaje       0.814       1.014       1.791
 Katus         -0.416       0.406       0.214
 Berendse       1.566       0.409       1.130
 Gorbacho       0.669       0.568       0.972
 Kiseljev      -0.792       0.273       0.351
 Gough         -0.232      -1.430       1.197
 Sherbati       0.552       0.776      -1.720
 Ghesquir       1.758      -0.198       0.974
 Avilov         0.486      -1.785      -0.219

 ... (loput jätetty tästä pois)
```

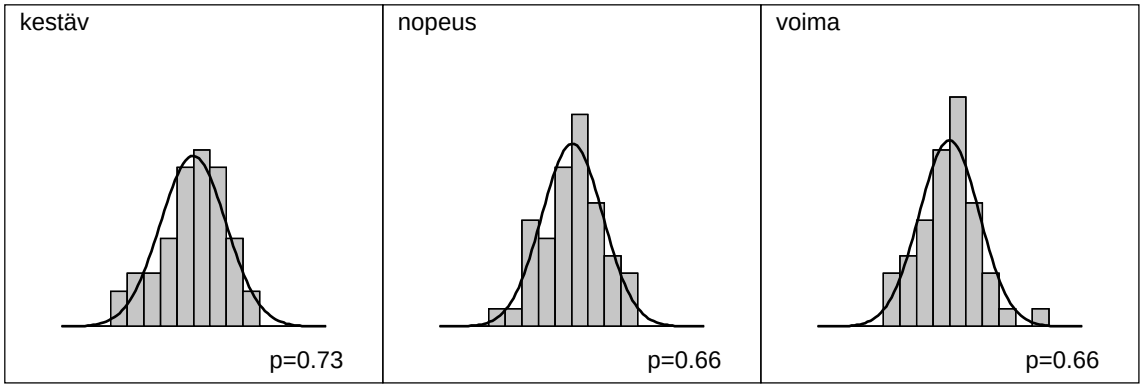
Piirretään havaintokohtaiset faktoripistemäärät pareittain vastakkain.



Kuten kuvistakin näkyy, faktoripistemäärät eivät korreloi keskenään. Tämä on jatkotarkasteluja silmälläpitäen hyödyllinen ominaisuus, esim. regressioanalyysi on mukavampaa korreloimattomilla selittäjillä. Keskiarvot ovat siis nollia ja hajonnat suunnilleen ykkösen suuruisia, eli faktoripistemäärät vastaavat melko tarkalleen standardoituja muuttujia.

```
CORR KYMMEN CUR+1 / VARS=kestäv,nopeus,voima
Means, std.devs and correlations of KYMMEN  N=48
Variable  Mean        Std.dev.
kestäv   -0.000000    0.983739
nopeus   -0.000000    0.941360
voima    -0.000000    0.896247
Correlations:
             kestäv  nopeus  voima
 kestäv       1.0000  0.0103 -0.0705
 nopeus       0.0103  1.0000 -0.0606
 voima       -0.0705 -0.0606  1.0000
```

Yleinen totuus on, että kun lasketaan yhteen erilaisia muuttujia, saadaan jotain enemmän tai vähemmän normaalijakaumaa muistuttavaa. Niinpä ei ole yllätys, että faktoripistemäärien jakaumat ovat selkeästi normaalisia, vaikkei havaintoja ole kuin 48.



Näin saatujen uusien faktoripistemuuttujien reliabiliteetit ovat varsin korkeita: 0.96, 0.88 ja 0.82. Näiden ja ao. muuttujien varianssien avulla voidaan laskea, että mittauksen keskivirheet ovat vastaavasti n. 0.2, 0.3 ja 0.4. Täten tiedetään miten tarkoista asteikoista nyt muodostetuissa faktoripisteissä on kysymys, ja esim. erilaisissa vertailutilanteissa voidaan arvioida, ylittääkö havaittu ero mittausvirheestä johtuvan vaihtelun.

Yleisesti käytössä oleva mitta, Cronbachin alfa, saisi jäädä jo historiaan. Tässä se antaa peräti negatiivisen tuloksen, mikä on täysin absurdia mitalle jonka pitäisi kuvata todellisen vaihtelun ja mittausvirhevaihtelun sisältämän kokonaisvaihtelun välistä suhdetta. *Lauri Tarkkosen* kehittämä mitta toimii kuten pitääkin, eli se ottaa huomioon ilmiön moniulotteisuuden. Mikäli käytettäisiin kolmen faktoripistemuuttujan sijasta suoraa summaa kaikista muuttujista, sen reliabiliteetti olisi vain luokkaa 0.26 eli todella huono. On tosin huomattava, että juuri yhteispisteitähän tässä lajissa käytännössä lasketaan, mutta tutkimuskäyttöön sellaisesta muuttujasta ei ole.

```
/RELIAB CORR.M,AFACT.M,MSN.M,3
Reliabilities of measurement scales by Tarkkonen's method, which
supersedes Cronbach's alpha (see RELIAB? for more information)

          Factor images        Factor scores
                E2     E3             E2     E3
            1   0.915  0.910      1   0.960  0.960
            2   0.818  0.820      2   0.884  0.884
            3   0.878  0.878      3   0.819  0.819

          Unweighted sum of all items
                E2     E3      Cronbach's alpha
                0.262  0.240   -0.483

E2: measurement errors are uncorrelated (assumed in factor model)
E3: measurement errors may correlate (more general model)
To test the assumptions of the model, see the residual matrices below:
    /MATSHOW RCOV.M  ##.### / Residual covariances
    /MATSHOW RCORR.M ##.### / Residual correlations
```

Estimoidun kolmen faktorin ratkaisun rakennevaliditeettia voidaan myös näiden tarkastelujen perusteella kyseenalaistaa. Puolella lajeista on alhainen kommunaliteetti, mikä näkyi itse asiassa jo faktorimatriisista:

```
                F3     F2     F1  Sumsqr
    Korkeus   0.126 -0.509  0.030  0.275 Korkeushyppy
    Pituush  -0.069  0.056 -0.272  0.082 Pituushyppy
    Aidat     0.060  0.220 -0.288  0.135 110 m aidat
    Seiväs   -0.255  0.020 -0.089  0.073 Seiväshyppy
    Keihäs    0.036 -0.266  0.016  0.072 Keihäänheitto
```

Kyseisille lajeille on yhteistä se että ne liittyvät tekniikkaan tai motoriikkaan. Tarkastellaan vielä residuaalien korrelaatiomatriisia:

```
Residual_correlations
///       100m Pituu Kuula Korke  400m Aidat Kiekk Seivä Keihä 1500m
100m      1.00 -0.04 -0.03 -0.02 -0.00 -0.03  0.03  0.01 -0.02 -0.00
Pituush  -0.04  1.00 -0.01  0.05  0.14  0.24  0.03  0.02  0.19  0.00
Kuula    -0.03 -0.01  1.00 -0.04  0.01  0.07  0.00 -0.00 -0.09 -0.00
Korkeus  -0.02  0.05 -0.04  1.00  0.02  0.09  0.04 -0.09  0.01  0.00
400m     -0.00  0.14  0.01  0.02  1.00  0.10 -0.01 -0.09  0.10  0.00
Aidat    -0.03  0.24  0.07  0.09  0.10  1.00 -0.07 -0.10 -0.10 -0.00
Kiekko    0.03  0.03  0.00  0.04 -0.01 -0.07  1.00 -0.00  0.11  0.00
Seiväs    0.01  0.02 -0.00 -0.09 -0.09 -0.10 -0.00  1.00 -0.12 -0.00
Keihäs   -0.02  0.19 -0.09  0.01  0.10 -0.10  0.11 -0.12  1.00  0.00
1500m    -0.00  0.00 -0.00  0.00  0.00 -0.00  0.00 -0.00  0.00  1.00
```

Faktorimallin mukaisesti tämän matriisin tulisi olla diagonaalinen, eli lävistäjän ulkopuolella pitäisi olla vain nollaa. Nyt residuaalien korrelaatiot osoittavat että osa vaihtelusta on jäänyt mittausvirheiden puolelle, ja mallia modifioimalla sitä voitaisiin siirtää todellisen vaihtelun puolelle. Käytännössä tämä tarkoittaisi faktorilukumäärän nostamista. Malliin tarvittaisiin yksi tai kaksi tekniikkafaktoria, käsien ja jalkojen taidoille erikseen.

Ns. *eksploratiivisen* faktorianalyysin puitteissa on sallittua toimia näin ja kehittää siten alkuperäistä konseptia moniulotteisemmaksi aineiston antaman informaation valossa. Tiukempi lähestymistapa eli ns. *konfirmatorinen* faktorianalyysi sen sijaan tarkoittaa faktorirakenteeseen liittyvien hypoteesien testaamista aiempaa tietämystä vasten. Eksploratiivinen työskentelytapa on käytännössä yleisemmin sovellettu. Faktorianalyysin yleistys useiden mittausmallien välisten suhteiden tutkimiseen tunnetaan puolestaan nimellä rakenneyhtälömallit *(structure equation models, SEM)*, josta käytetään myös usein nimitystä LISREL-mallit (samannimisen ohjelmiston perusteella).

Tietynlaista konfirmatorista työskentelytapaa edustaa myös *transformaatioanalyysi*, jolla voidaan vertailla faktorirakenteita toisiinsa, esim. eri tutkimusten, ajankohtien tms. välillä. Suoraan vertailuja ei pidä mennä tekemään, sillä rotaatiosta johtuen identtisetkin rakenteet voivat näyttää erilaisilta. Transformaatioanalyysin kehitti alunperin *Yrjö Ahmavaara* jo 1950-luvulla. *Seppo Mustonen* laati siitä myöhemmin ns. symmetrisen mallin. Kansainvälisesti menetelmä ei ole kovin tunnettu. Lähimmäksi sitä tulee ns. *Procrustes*-rotaatio, mutta siinä ei siten kuin transformaatioanalyysissa kiinnitetä huomiota poikkeamiin sen jälkeen kun ratkaisut lähimmäksi tuova rotaatio on löydetty. Juuri poikkeamat ovat kuitenkin mielenkiintoisia, sillä ne kertovat mistä mahdolliset rakenne-erot johtuvat. Mm. kulttuurierot kyselytutkimuksissa, joissa lomake on huolellisesti käännätetty toiselle kielelle, paljastuvat armotta.

Edellä esitetty esimerkki sopii faktorianalyysin periaatteiden esittämiseen, mutta todellisuudessa aineiston pitäisi olla kooltaan suurempi. Hyötysuhdekin jää vaatimattomaksi, jos alunperin kymmenestä muuttujasta saadaan ulotteisuus tiivistettyä viiteen. Myös havaintojen suhteen aineisto on kovin pieni. Suurimman uskottavuuden estimointi on vakaammalla pohjalla, kun estimoitavia parametreja kohti on enemmän havaintoja. Tyypillisempiä aineistokokoja faktorianalyysissa ovatkin sellaiset, joissa muuttujia on 30-150 ja havaintoja 100-1000. Yleisiä suosituksia on mahdotonta antaa. Esim. mittaustarkkuus vaikuttaa asiaan: mitä karkeammilla mittareilla mitataan, sitä enemmän olisi oltava havaintoja. Pienemmilläkin aineistoilla voidaan toimia, mutta tulokset jäävät helposti hatariksi. Havaintoja on joka tapauksessa oltava enemmän kuin muuttujia. Tämä pätee moniin muihinkin menetelmiin.

SAS:issa faktorianalyysi tehdään proseduurilla FACTOR. Faktorointimenetelmä pitää muistaa valita eksplisiittisesti, sillä oletuksena tehdään pääkomponenttianalyysi. Asiallisia reliabiliteettitarkasteluja ei vielä ole SAS:issa yleisessä käytössä, mutta ne lienee verrattain helppo ohjelmoida makroina esim. SAS:in matriisikielellä (IML).

Lisää tietoa mittauksen mallintamisesta ja reliabiliteetin arvioinnista saa vuonna 2000 julkaistusta väitöskirjastani, joka on luettavissa verkossa osoitteessa *http://ethesis.helsinki.fi/julkaisut/val/tilas/vk/vehkalahti/*

Pääkomponenttianalyysi

Pääkomponenttianalyysi *(principal components analysis, PCA)* on yksinkertainen keino tiivistää keskenään korreloivien muuttujien sisältämää informaatiota uusiksi, korreloimattomiksi muuttujiksi.

Esimerkkinä on aineisto, jossa on tietoja kahdesta myyrälajista, *microtus multiplex* ja *microtus subterraneus*. Lajit on morfometrisesti vaikea erottaa toisistaan. Muuttujat ovat erilaisia näistä tehtyjä mittauksia. Ensimmäistä lajia on havaittu 43 ja toista 46 kpl. Loppuja N-(43+46)=199 ei ole saatu tunnistettua riittävän aukottomasti. Voidaan osoittaa että erotteluanalyysin avulla erot tulevat varsin selvästi esille, ja että tunnistamattomat yksilöt voidaan luotettavasti luokitella oikeaan ryhmään.

Tarkastellaan tässä yhteydessä kuitenkin vain jälkimmäistä lajia, josta on siis käytettävissä 46 havaintoa. Muuttujien kuvaukset käyvät ilmi aineistosta.

```
FILE STATUS MICROTUS
Flury, B. (1997). A First Course in Multivariate Statistics.
FIELDS: (active)
   1 NA_   1 Group    1=M.multiplex, 2=M.subterraneus, 0=not determined
   2 NA_   2 M1LEFT   Width of upper left molar 1
   3 NA_   2 M2LEFT   Width of upper left molar 2
   4 NA_   2 M3LEFT   Width of upper left molar 3
   5 NA_   2 FORAMEN  Length of incisive foramen
   6 NA_   2 PBONE    Length of palatal bone
   7 NA_   2 LENGTH   Condylo incisive length or skull length
   8 NA_   2 HEIGHT   Skull height above bullae
   9 NA_   2 ROSTRUM  Skull width across rostrum
END
Survo data file MICROTUS: record=41 bytes, M1=15 L=64  M=9 N=288
```

Valitaan muuttujat (kaikki paitsi ryhmätunnus) ja lasketaan otossuureet lajin 2 osalta.

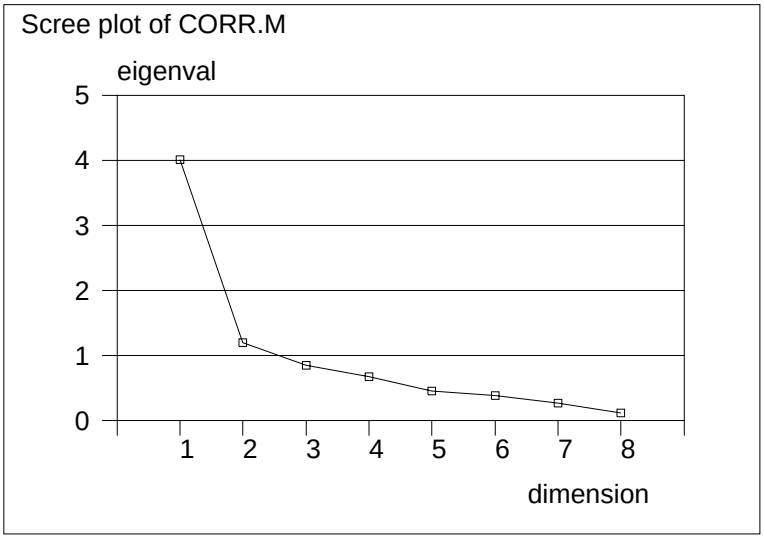
```
MASK=-AAAAAAAA
CORR MICROTUS / IND=Group,2
```

Muuttujat mittaavat paljolti samoja asioita, mikä on aivan tavallista tämäntyyppisissä aineistoissa.

```
/LOADMSN MICROTUS
Means, standard deviations and number of observations in MICROTUS:
            mean   stddev    N
M1LEFT   1773.261   62.526   46 Width of upper left molar 1
M2LEFT   1504.630   63.331   46 Width of upper left molar 2
M3LEFT   1597.804  125.969   46 Width of upper left molar 3
FORAMEN  3899.043  200.538   46 Length of incisive foramen
PBONE    4805.217  274.945   46 Length of palatal bone
LENGTH   2226.674   93.465   46 Condylo incisive length or skull length
HEIGHT    758.065   22.673   46 Skull height above bullae
ROSTRUM   427.196   17.871   46 Skull width across rostrum
```

```
/LOADCORR
Limits: P=0.001 0.469   P=0.01 0.372   P=0.05 0.29
         M1LEFT M2LEFT M3LEFT FORAME  PBONE LENGTH HEIGHT ROSTRU
M1LEFT    1.000  0.228  0.430  0.434  0.266  0.546  0.246  0.377
M2LEFT    0.228  1.000  0.416  0.517  0.380  0.615  0.197  0.549
M3LEFT    0.430  0.416  1.000  0.425  0.234  0.690  0.449  0.494
FORAMEN   0.434  0.517  0.425  1.000 -0.014  0.695  0.114  0.505
PBONE     0.266  0.380  0.234 -0.014  1.000  0.429  0.431  0.416
LENGTH    0.546  0.615  0.690  0.695  0.429  1.000  0.447  0.729
HEIGHT    0.246  0.197  0.449  0.114  0.431  0.447  1.000  0.390
ROSTRUM   0.377  0.549  0.494  0.505  0.416  0.729  0.390  1.000
```

Piirretään pääkomponenttianalyysiin liittyvä Scree-kuva, joka havainnollistaa korrelaatiomatriisin todellisia ulottuvuuksia kuvaavia ominaisarvoja *(eigenvalues)* järjestyksessä suurimmasta pienimpään.



Pääkomponentteja voidaan muodostaa saman verran kuin muuttujia, mutta koska tavoitteena on aineiston tiivistäminen, on haettava sopiva katkaisukohta. Koska ominaisarvot vastaavat pääkomponenttien variansseja, on tavallista katsoa esim. kuvasta, milloin ominaisarvot putoavat selvästi alle yhden. Tämä on luonnollinen katkaisukohta: yksittäisten standardoitujen muuttujien varianssit ovat ykkösiä, joten on yleensä turha ottaa mukaan pääkomponenttia jolla on pienempi varianssi kuin yksittäisellä muuttujalla. Näin voidaan kuitenkin tehdä mikäli tälle pääkomponentille löytyy järkevä tulkinta.

Tässä kohtaa on huomautettava, että em. toimintatapaa näkee usein sovellettavan virheellisesti myös faktorianalyysin yhteydessä. Analyysit muistuttavatkin toisiaan päällisin puolin, mutta niillä on perusteellinen ero: faktorianalyysi perustuu tilastolliseen malliin, kun taas pääkomponenttianalyysi on pelkkä laskennallinen apuväline. Pääkomponenttianalyysille kelpaa kaikki vaihtelu sellaisenaan, mutta faktorianalyysi kykenee (malliperustansa ansiosta) erottelemaan todellisen vaihtelun virhevaihtelusta.

Molemmilla menetelmillä on käyttöalueensa. Jos tutkimusasetelmassa on luonnollista ajatella piilomuuttujia, on faktorianalyysi oikea valinta. Erityisesti näin on kysely- ja haastattelututkimusten yhteydessä, mutta myös mm. aistinvaraiseen arviointiin perustuvissa tutkimusasetelmissa. Pääkomponenttianalyysi puolestaan soveltuu tilanteisiin, joissa halutaan vain suoraviivaisesti kasata useita (korreloituneita) muuttujia yhteen. Tyypillisiä esimerkkejä ovat ekologiset aineistot, joissa saattaa olla kymmeniä samantyyppisiä vahvasti keskenään korreloivia mittauksia.

Valitettavasti faktorianalyysin pitkän kehityskulun aikana asiat ovat päässeet sekoittumaan, mistä johtuen nämä kaksi menetelmää usein sotketaan toisiinsa, tavalla tai toisella. Asiaa ei ainakaan yhtään selvennä se, että mm. SAS:issa ja SPSS:ssä faktori- ja pääkomponenttianalyysi suoritetaan samalla proseduurilla, ja *oletuksena tehdään pääkomponenttianalyysi*.

Lasketaan nyt kaikki 8 pääkomponenttia myyräaineistosta.

```
/PCOMPR CORR.M,MSN.M,8
```

Pääkomponenttimatriisi sisältää pääkomponenttien ja alkuperäisten muuttujien väliset korrelaatiot. Sen pohjalta voidaan pääkomponenteille löytää tulkinnat. Usein ensimmäinen komponentti kerää valtaosan muuttujien varianssista, tässäkin tapauksessa puolet.

```
Principal_components
///       PCOMP1  PCOMP2  PCOMP3  PCOMP4  PCOMP5  PCOMP6  PCOMP7  PCOMP8
M1LEFT    -0.619   0.120  -0.503   0.555   0.033  -0.068   0.187  -0.005
M2LEFT    -0.706   0.133   0.558  -0.028   0.221  -0.241   0.254   0.012
M3LEFT    -0.750   0.017  -0.297  -0.324   0.445   0.186  -0.045  -0.093
FORAMEN   -0.684   0.611   0.015  -0.038  -0.209  -0.198  -0.223  -0.155
PBONE     -0.522  -0.673   0.274   0.363   0.080  -0.005  -0.231  -0.090
LENGTH    -0.935   0.095  -0.009  -0.020  -0.020   0.027  -0.184   0.284
HEIGHT    -0.551  -0.584  -0.292  -0.362  -0.235  -0.275   0.093  -0.011
ROSTRUM   -0.811  -0.019   0.193  -0.032  -0.320   0.414   0.163  -0.053

Variances_of_principal_components
///       PCOMP1  PCOMP2  PCOMP3  PCOMP4  PCOMP5  PCOMP6  PCOMP7  PCOMP8
Variance   4.022   1.209   0.851   0.680   0.456   0.384   0.273   0.124

Variances_of_pr.components_(in_percentages)
///       PCOMP1  PCOMP2  PCOMP3  PCOMP4  PCOMP5  PCOMP6  PCOMP7  PCOMP8
Per_cent  50.275  15.110  10.639   8.500   5.701   4.804   3.418   1.552
Cumulat.  50.275  65.385  76.024  84.524  90.225  95.029  98.448 100.000
```

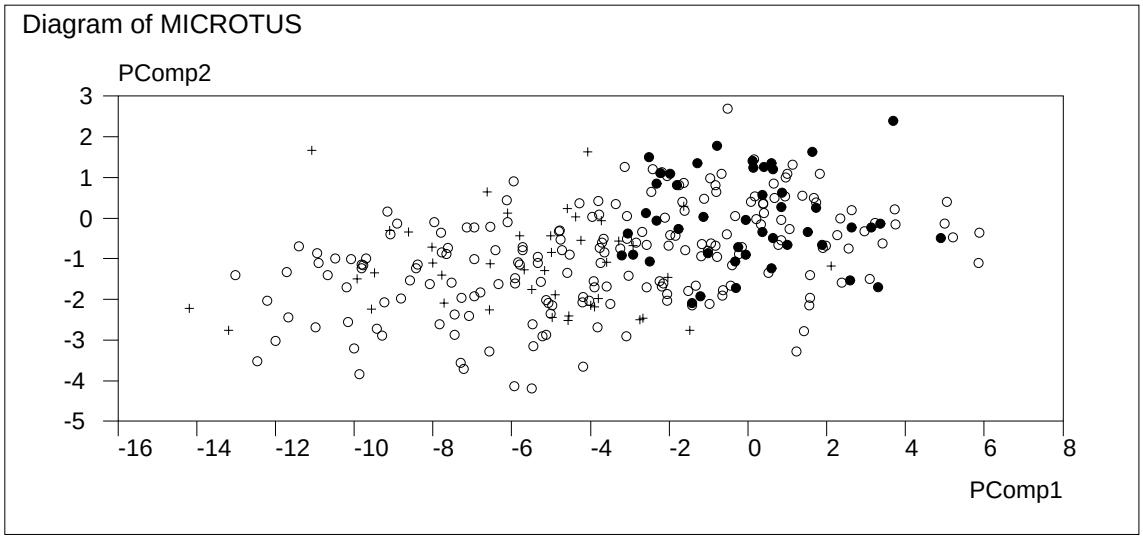
Lasketaan havaintokohtaiset pääkomponenttipistemäärät painotettuina summina alkuperäisistä muuttujista. Tyydytään kahteen ensimmäiseen pääkomponenttiin, jotka selittävät 65.4 % vaihtelusta. Vaikka analyysi tehtiinkin vain lajin 2 havainnoilla, lasketaan pistemäärät myös muille havainnoille.

```
LINCO MICROTUS,PCOEFF.M(PComp1,PComp2)
```

Näin saatuja uusia muuttujia voidaan käyttää missä tahansa jatkoanalyyseissa. Ideana on että ne korvaavat alkuperäiset kahdeksan muuttujaa, jolloin aineisto on saatu tiiviimmäksi. Pääkomponenttipistemäärät ovat korreloimattomia, joten ne ovat mukavampia kuin alkuperäiset korreloivat muuttujat. Tämän kaiken kustannus on se että 100-65.4=34.6 % vaihtelusta on jäänyt tarkastelujen ulkopuolelle. Lisäksi olisi tärkeää saada edellä olevien matriisien avulla haettua kunnolliset tulkinnat uusille muuttujille. Ekologian taitoni eivät riitä tämän esimerkin syvällisempään pohdintaan, joten on tyydyttävä hieman pinnallisempaan esittelyyn.

Usein on hyödyllistä piirtää esimerkiksi pääkomponentteja vastakkain.

```
GPLOT MICROTUS,PComp1,PComp2 / POINT=0,4,Group,0
```



Tässä korreloimattomuus koskee vain lajia 2, jonka perusteella pääkomponentit laskettiin. Sitä vastaavat pisteet erottuvat kuvassa värjättyinä. Koko aineiston osalta pääkomponenttipistemäärät korreloivat kuvan perusteella selvästi. Lajia 1 edustavat +:lla ja tunnistamattomia o:lla merkityt pisteet.

Eräät muut monimuuttujamenetelmät kuten moniulotteinen skaalaus ja korrespondenssianalyysi ovat läheistä sukua pääkomponenttianalyysille. Niiden yhteydessä tämänkaltaiset graafiset tarkastelut korostuvat vielä enemmän.

SAS:issa pääkomponenttianalyysin voi siis tehdä FACTOR-proseduurilla, mutta sille on omakin proseduurinsa, PRINCOMP.

Erotteluanalyysi

Erotteluanalyysissa *(discriminant analysis)* haetaan vastausta siihen, mikä parhaiten erottaa (tunnetut) ryhmät toisistaan. Lisäksi analyysiin liittyy usein kiinteästi luokittelu *(classification)*, jossa selvitetään, mihin ryhmään (mahdollinen uusi) havainto kuuluu.

Menetelmässä haetaan sellaisia muuttujien yhdistelmiä, joilla ryhmien väliset erot tulevat mahdollisimman suuriksi. Näin muodostettavia lineaarikombinaatioita kutsutaan erottelumuuttujiksi tai -funktioiksi *(discriminant function)*.

Tarkastellaan Suomen alueellisia eroja 1980-luvun alun kunta-aineiston valossa. Katsotaan mitkä tekijät erottavat eri maantieteellisiä alueita.

```
FILE STATUS KUNNAT
  Suomen kunnat aakkosjärjestyksessä
  Tiedot ovat pääosin vuosilta 1978-80.      5.2.84/SM
FIELDS: (active)
   1 SA-  16 Kunta    Kunnan nimi
   2 SA-   3 Lääni    UUS,TUR,AHV,HÄM,KYM,MIK,KAR,KUO,KES,VAA,OUL,LAP
   3 NA-   4 Väestö   Arvioitu maassa asuva väestö 1.1.1980 (######)
   4 NA-   4 Synt.    Elävänä syntyneet v.1978 (####)
   5 NA-   8 Ala      Maapinta-ala km^2 1.1.1979 (#####.##)
   6 SA-   1 Maamet   Maa- ja metsätaloudessa toimivien osuus (10%)
   7 SA-   1 Teoll    Teollisuudessa toimivien osuus (10%)
   8 SA-   1 Palvelu  Ammatissa ja palveluelinkeinoissa toim.osuus (10%)
   9 NA-   4 Asuin    Valmistuneet asuinhuoneistot v.1978 (####)
  10 NA-   4 Äyri     Veroäyrin hinta v.1979 (##.##)
  11 NA-   2 Tulotaso Veroäyrejä asukasta kohti v.1979 (#####)
  12 NA-   4 SYNT     1000*Synt./Väestö (##.###)
END
Survo data file KUNNAT: record=128 bytes, M1=30 L=64  M=14 N=464
```

Määritellään aineistoon neljä elinkeinoaluetta vanhan läänijaon perusteella. Tyypillisesti tätä vaihetta ei tietenkään tarvita vaan aineistossa on yleensä valmiina jokin luokitteluasteikollinen muuttuja, johon ryhmäjako perustuu.

```
CLASSIFY KUNNAT,ALUEET,Lääni,Alue
CLASSIFICATION ALUEET
UUS,TUR,HÄM,KYM:      Ruuhka
MIK,KAR,KUO,KES,VAA:  Keski
OUL,LAP:              Pohjois
AHV:                  Saaristo
END
```

Valitaan muuttujat joiden perusteella erottelu tehdään ja lasketaan ryhmittäiset otossuureet. Talletetaan ne omiin matriisitiedostoihinsa.

```
MASK=----AAAAAAA----
CORR KUNNAT / CASES=Alue:Ruuhka
MAT M1=MSN.M
MAT R1=CORR.M
CORR KUNNAT / CASES=Alue:Keski
MAT M2=MSN.M
MAT R2=CORR.M
CORR KUNNAT / CASES=Alue:Pohjois
MAT M3=MSN.M
MAT R3=CORR.M
CORR KUNNAT / CASES=Alue:Saaristo
MAT M4=MSN.M
MAT R4=CORR.M
```

Poimitaan ryhmittäisiä tietoja esille matriiseista M1-M4.

Ruuhka	mean	stddev	N	**Keski**	mean	stddev	N
Ala	287.280	221.661	213	Ala	581.905	433.893	161
Maamet	2.451	1.884	213	Maamet	3.565	1.706	161
Teoll	3.052	1.350	213	Teoll	2.130	1.168	161
Palvelu	3.023	1.195	213	Palvelu	2.907	1.054	161
Asuin	153.404	408.826	213	Asuin	89.447	128.869	161
Äyri	15.646	0.699	213	Äyri	16.801	0.654	161
Tulotaso	14113.676	3311.226	213	Tulotaso	11597.832	2331.916	161

Pohjois	mean	stddev	N	**Saaristo**	mean	stddev	N
Ala	2035.663	2815.727	74	Ala	92.579	40.438	16
Maamet	3.081	1.710	74	Maamet	3.000	1.506	16
Teoll	2.041	1.103	74	Teoll	1.063	0.854	16
Palvelu	3.419	1.293	74	Palvelu	4.375	1.258	16
Asuin	106.081	159.497	74	Asuin	19.750	26.514	16
Äyri	17.044	0.636	74	Äyri	14.344	0.724	16
Tulotaso	11192.135	2293.994	74	Tulotaso	13041.125	2421.540	16

Näihin sekä ryhmittäisiin korrelaatioihin pohjautuen tehdään erotteluanalyysi.

```
 MSN=M1,M2,M3,M4 (keskiarvojen ja hajontojen matriisit)
CORR=R1,R2,R3,R4 (korrelaatiomatriisit)
/DISCRI
    Eig.val.  %      Can.corr  Chi^2    df  P
 1  1.258817  68.80  0.746519  596.5672  21 1
 2  0.431721  23.60  0.549126  223.7773  12 1
 3  0.139117   7.60  0.349467   59.5910   5 0.9999
```

Tulostuksesta nähdään että ensimmäinen erottelumuuttuja selittää lähes 70 % ryhmien välisestä vaihtelusta. Seuraava, ensimmäiseen nähden ortogonaalinen erottelumuuttuja selittää n. 24 %. Loppuosa vaihtelusta jää kolmannelle dimensiolle, joka vastaa enää 8 % osuudesta kokonaisvaihtelusta. Kaikki erottelumuuttujat ovat tilastollisesti merkitseviä, mikä johtuu osaltaan melko suuresta havaintojen määrästä. Käytännössä kolmas erottelumuuttuja ei ole kovin mielenkiintoinen, joten riittää tarkastella pääasiassa kahta ensimmäistä.

Tulkinnan kannalta keskeisessä asemassa ovat erottelumuuttujien ja alkuperäisten muuttujien väliset korrelaatiot, eli ns. rakennematriisi.

```
Correlations_between_variables_and_discriminators
///      Discr1 Discr2 Discr3
Ala      -0.498  0.267  0.710 Maapinta-ala km^2 1.1.1979 (#####.##)
Maamet   -0.268  0.105 -0.493 Maa- ja metsätaloudessa toimivien osuus (10%)
Teoll     0.298 -0.544  0.430 Teollisuudessa toimivien osuus (10%)
Palvelu   0.046  0.423  0.219 Ammatissa ja palveluelinkeinoissa toim.osuus (10%)
Asuin     0.072 -0.148  0.190 Valmistuneet asuinhuoneistot v.1978 (####)
Äyri     -0.958 -0.164 -0.201 Veroäyrin hinta v.1979 (##.##)
Tulotaso  0.514 -0.247  0.311 Veroäyrejä asukasta kohti v.1979 (#####)
```

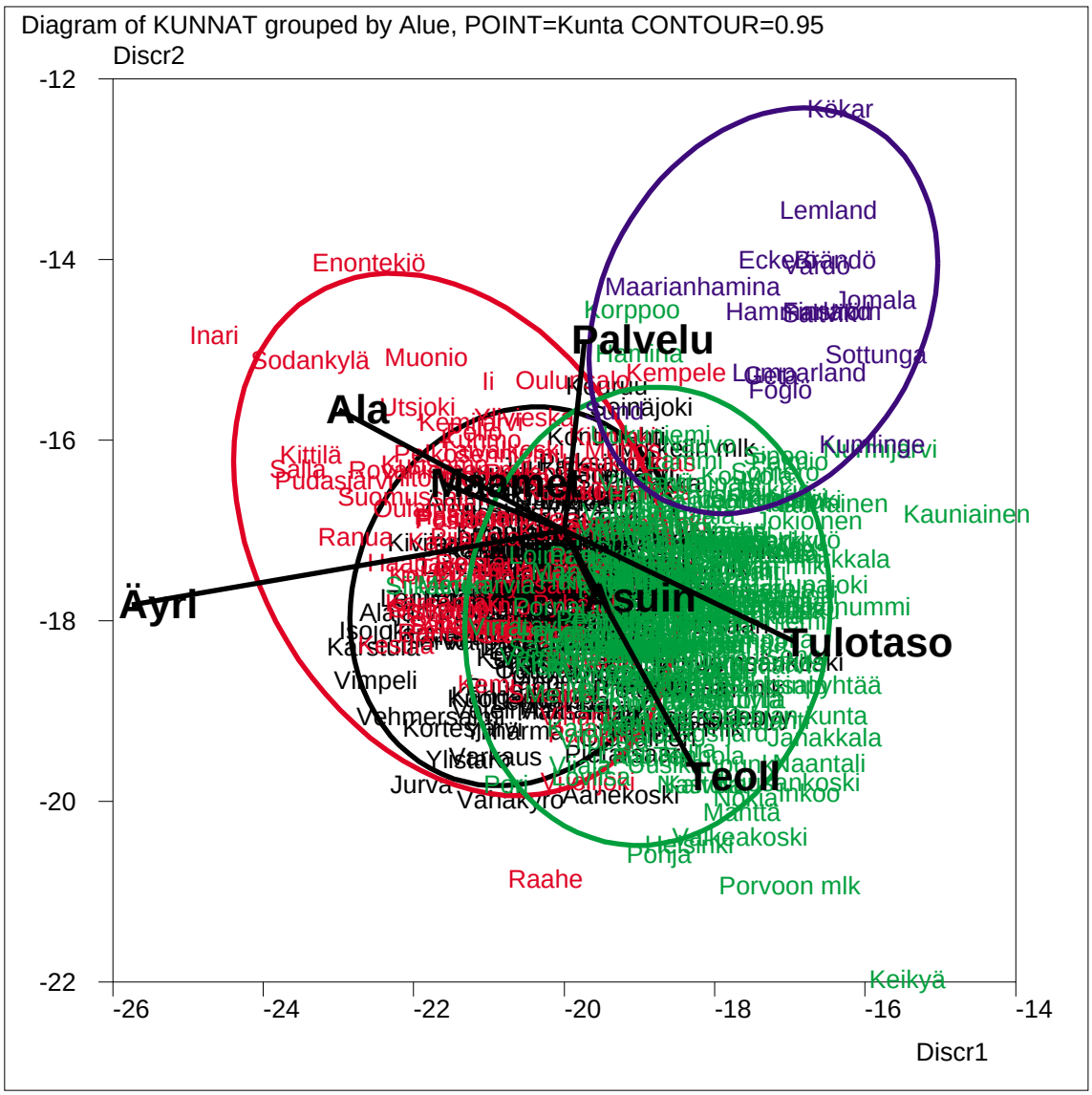
Korrelaatioista nähdään että ensimmäinen erottelumuuttuja perustuu paljolti kunnan vaurauteen. Toisessa korostuvat eniten teollisuus- ja palveluelinkeinoissa toimivien osuudet, erisuuntaisina ääripäinä.

Lasketaan havaintokohtaiset erottelupistemäärät painotettuina summina alkuperäisistä muuttujista.

```
LINCO KUNNAT,DISCRL.M(Discr1,Discr2)
```

Erottelumuuttujien rajaama ns. erotteluavaruus kannattaa havainnollistaa graafisesti piirtämällä dimensiot vastakkain ja kunkin ryhmän pisteet toisistaan erottuvin merkein (tässä on käytetty kunnan nimeä). Ryhmien ympärille on piirretty hajontaellipsit, jotka auttavat hahmottamaan paremmin ryhmien olemusta. Tässä ne on piirretty siten että 95 % ryhmän havainnoista on ellipsin sisäpuolella. Lisäksi kuvassa näkyvät erottelun perustana olevat muuttujat origosta lähtevinä vektoreina. Ne kuvaavat edellä olevan rakennematriisin tietoja, jolloin samasta kuvasta nähdään lopulta kaikki keskeiset tiedot yhtaikaa.

```
MASK=----AAAAAAA----XY
/DCONTOUR KUNNAT Alue / POINT=Kunta
```



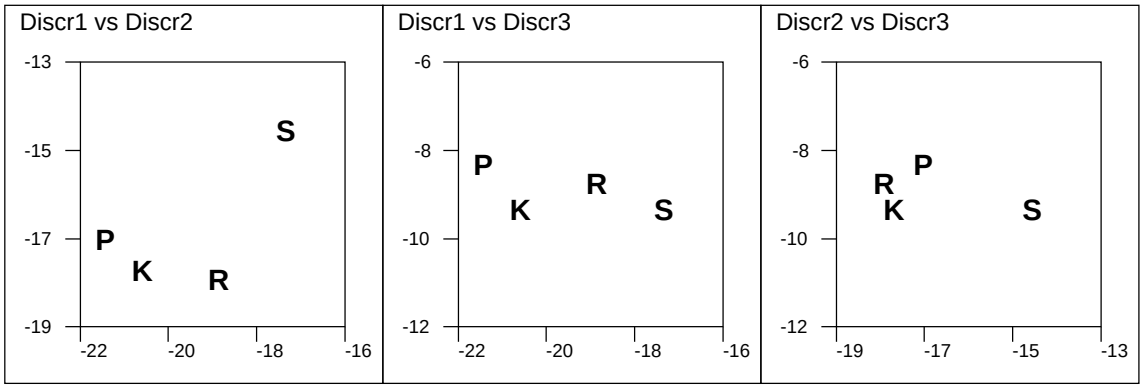
Alueet etelästä pohjoiseen ryhmittyvät melko selvästi jo ensimmäisen erottelumuuttujan mukaan, mutta saariston kuntien toisenlainen elinkeinorakenne vaatii myös toisen erottelumuuttujan. Vaurauden suhteen Ruuhka-Suomen ja Saaristo-Suomen kunnat eivät paljoa eroa toisistaan.

Kun havaintopisteen paikalle laitetaan havainnon tunnus (kuten tässä kunnan nimi), ei ole tarkoituskaan saada selvää mitä ryhmien ytimissä tapahtuu. Mielenkiinto kohdistuu ääritapauksiin kuten esimerkiksi Inari, Enontekiö, Kökar, Kauniainen, Keikyä ja Raahe.

Kuvan perusteella saattaa näyttää siltä, etteivät ryhmät juuri eroaisi toisistaan. Erot ovat kuitenkin selvempiä mitä luulisi. Sen voi todeta esim. erottelumuuttujien ryhmäkeskiarvojen avulla.

```
Groupwise_means_of_discriminators
///      Discr1 Discr2 Discr3
Ruuhka   -18.89 -17.95  -8.75
Keski    -20.61 -17.73  -9.35
Pohjois  -21.45 -17.04  -8.33
Saaristo -17.35 -14.56  -9.35
```

Piirretään keskiarvot erotteluavaruuteen eri suunnilta:



Ensimmäisellä dimensiolla kaikki erot ovat selviä. Toisellakin vain Ruuhka- ja Keski-Suomen ero jää mitättömäksi. Saman asian voi halutessaan todentaa analyyttisesti varianssianalyysin avulla.

Tehdään yksisuuntainen varianssianalyysi ja parivertailut toiselle erottelumuuttujalle.

```
ANOVA KUNNAT,END+2 / DEPENDENT=Discr2 GROUPING=Alue METHOD=ONEWAY,TUKEY(S)

Results for dependent variable Discr2:
Means and deviations

                Keski       Ruuhka      Pohjois     Saaristo       Total
Means         -17.72585   -17.94624   -17.04157   -14.56410   -17.60886
Deviations      0.85734     1.03694     1.18107     0.91702     1.19266
N of obs.           161         213          74          16         464

One-way fixed effects analysis of variance
Source          sum of squares        df     mean squares
Between groups     198.591626962       3    66.1972089874
Within groups      460.000004588     460    1.00000000997
Total              658.591631551     463

F test for equality of means: F-value =  66.1972
It equals the  100.0% point of the F(3, 460) distribution.
Risk of rejecting the nullhypothesis, when true, is 0.000000


Test for equality of means without assuming equal group variances:
Brown-Forsythe statistic =  65.1969 with df 3 and   140.98
Appr. risk of rejecting the null hypothesis, when true, is 0.000000
```

```
Multiple comparisons of means by the Tukey-Kramer method

Pairwise mean differences

Degrees of freedoms for each test statistic are 4 and 460

For each test statistic the lowest experimentwise significance level
at which the null hypothesis can be rejected are given

The mean of the first set of group means will be compared with
the mean of the second set of group means

First set  Second set  contrast test st.   prob.

Keski      Ruuhka       0.22039  2.98452  0.1519
Keski      Pohjois     -0.68428 -6.89039  0.0000
Keski      Saaristo    -3.16175 -17.0580  0.0000
Ruuhka     Pohjois     -0.90467 -9.48135  0.0000
Ruuhka     Saaristo    -3.38214 -18.4518  0.0000
Pohjois    Saaristo    -2.47747 -12.7080  0.0000
```

Kun erottelu on tehty, kuvannettu ja tulkittu, on aika siirtyä luokitteluongelman pariin. Kysymys kuuluu siis: "Mihin ryhmään havainto kuuluu?"

Luokitteluperusteita on useampia. Eräs niistä perustuu havaintojen välisiin Mahalanobis-etäisyyksiin. Kyseessä on monimuuttujamenetelmien yhteydessä usein esiintyvä etäisyysmitta, joka ottaa huomioon muuttujien väliset korrelaatiot. Toinen mahdollisuus on käyttää Bayes-periaatetta, eli oletetaan ryhmiin kuulumisille tietyt a priori -todennäköisyydet ja sijoitetaan havainto siihen ryhmään, jolle multinormaalijakauman tilanteessa a posteriori -todennäköisyyteen verrannollinen lauseke on suurin. Kummankin periaatteen osalta voidaan vielä luokittelu tehdä ryhmien omien tai yhteisten *(pooled)* kovarianssirakenteiden perusteella.

Sovelletaan kaikkia näitä tapoja. Lasketaan myös Bayes-periaatteen (ryhmittäiset kovarianssit) mukaiset posterioritodennäköisyydet kuhunkin ryhmään kuulumiselle. Lisätään aineistoon muutamia uusia muuttujia havaintojen luokittelemiseksi.

```
FILE UPDATE KUNNAT
FIELDS:
  18 ND-   1 Mahal1    Mahalanobis-etäisyys, samat kovarianssit
  19 Nd-   1 Mahal2    Mahalanobis-etäisyys, ryhmittäiset kovarianssit
  20 NB-   1 Bayes1    Bayes-todennäköisyys, samat kovarianssit
  21 Nb-   1 Bayes2    Bayes-todennäköisyys, ryhmittäiset kovarianssit
  22 NP-   4 P1        Posterioritodennäköisyys (Ruuhka)
  23 NP-   4 P2        Posterioritodennäköisyys (Keski)
  24 NP-   4 P3        Posterioritodennäköisyys (Pohjois)
  25 NP-   4 P4        Posterioritodennäköisyys (Saaristo)
END
```

Luokitellaan kunnat erottelumuuttujien avulla.

```
 MSN=DM1,DM2,DM3,DM4 (erottelumuuttujien keskiarvojen ja hajontojen matriisit)
CORR=DR1,DR2,DR3,DR4 (erottelumuuttujien korrelaatiomatriisit)
CLASSI KUNNAT / COEFF=DISCRL.M
```

Tarkistetaan luokittelun onnistuminen ristiintaulukoimalla luokittelut alkuperäistä ryhmäjakoa vastaan. Lasketaan kunkin luokitteluperiaatteen osalta oikein luokiteltujen kuntien osuus ja tutkitaan hieman virheellisesti luokiteltuja tapauksia.

```
    Alue=/Ruuhka,/Keski,/Pohjois,/Saaristo
 *luokat=1,1(Ruuhka),2(Keski),3(Pohjois),4(Saaristo)
  Mahal1=*luokat Bayes1=*luokat
  Mahal2=*luokat Bayes2=*luokat          N=464

TAB KUNNAT,END+3 / VARIABLES=Mahal1,Alue CHI2=- LABELS=0
TAB KUNNAT,END+3 / VARIABLES=Mahal2,Alue
TAB KUNNAT,END+3 / VARIABLES=Bayes1,Alue
TAB KUNNAT,END+3 / VARIABLES=Bayes2,Alue

          Mahal1   Ruuhka    Keski  Pohjois Saaristo
Alue      ******
/Ruuhka               166       39        0        8
/Keski                 30      116       15        0
/Pohjois                5       29       39        1
/Saaristo               0        1        0       15
            336/N=0.72

          Mahal2   Ruuhka    Keski  Pohjois Saaristo
Alue      ******
/Ruuhka               172       16       20        5
/Keski                 29       48       84        0
/Pohjois                6        8       59        1
/Saaristo               0        0        1       15
            294/N=0.63

          Bayes1   Ruuhka    Keski  Pohjois Saaristo
Alue      ******
/Ruuhka               182       28        0        3
/Keski                 33      124        4        0
/Pohjois               10       35       29        0
/Saaristo               0        1        0       15
            350/N=0.75

          Bayes2   Ruuhka    Keski  Pohjois Saaristo
Alue      ******
/Ruuhka               182       30        0        1
/Keski                 33      125        3        0
/Pohjois                7       38       29        0
/Saaristo               0        1        0       15
            351/N=0.76
```

Parhaimmillaan päästään 76 prosenttiin. Suurimmat virheet tapahtuvat Ruuhka- ja Keski-Suomen kuntien luokittelussa. Huonoin tulos saadaan Mahalanobis-etäisyyksiin perustuvalla luokittelulla jossa käytetään ryhmittäisiä kovariansseja: merkittävä osa Keski-Suomen kunnista on luokiteltu kuuluvaksi Pohjois-Suomeen. Usein onkin suositeltavampaa olettaa kovarianssit samoiksi.

Tässä on huomattava, että luokittelun tulos on liian hyvä, koska se tehtiin samasta aineistosta kuin erottelu. Paremman kuvan luokittelun onnistumisesta saa jakamalla aineiston kahtia, tekemällä erottelun toiseen osaan ja luokittelemalla sillä perusteella toisen osan. Havaintoja olisi siis suotavaa olla runsaasti.

Tarkastellaan vielä posterioritodennäköisyyksien jakaumia.

```
STAT KUNNAT END+2 / VARS=P1,P2,P3,P4

Basic statistics: KUNNAT N=464
Variable: P1        Posterioritodennäköisyys (Ruuhka)
min=0        in obs.#19 (Enontekiö)
max=1        in obs.#117 (Keikyä)
mean=0.466539 stddev=0.372705 skewness=0.13997  kurtosis=-1.601591
lower_Q=0.092568 median=0.420455 upper_Q=0.864773
up.limit      f     %      *=2 obs.  class width=0.05
    0         8   1.7 ****
    0.05     77  16.6 ************************************
    0.1      37   8.0 ******************
    0.15     34   7.3 ****************
    0.2      19   4.1 *********
    0.25     12   2.6 ******
    0.3      16   3.4 ********
    0.35     13   2.8 ******
    0.4      12   2.6 ******
    0.45     11   2.4 *****
    0.5      11   2.4 *****
    0.55      9   1.9 ****
    0.6      12   2.6 ******
    0.65     16   3.4 ********
    0.7       9   1.9 ****
    0.75     19   4.1 *********
    0.8      10   2.2 *****
    0.85     17   3.7 ********
    0.9      22   4.7 ***********
    0.95     34   7.3 ****************
    1        66  14.2 *********************************
```

Osa on luokiteltu aivan selvästi Ruuhka-Suomeen kuuluviksi, osa puolestaan yhtä selvästi sinne kuulumattomiksi. Epäselviä tapauksiakin riittää.

```
Variable: P2        Posterioritodennäköisyys (Keski)
min=0        in obs.#19 (Enontekiö)
max=0.886496 in obs.#394 (Tervo)
mean=0.384451 stddev=0.320205 skewness=0.180336 kurtosis=-1.560897
lower_Q=0.052206 median=0.348214 upper_Q=0.720536
up.limit      f     %      *=2 obs.  class width=0.05
    0.05    115  24.8 ************************************************************
    0.1      34   7.3 *****************
    0.15     20   4.3 **********
    0.2      14   3.0 *******
    0.25     18   3.9 *********
    0.3      18   3.9 *********
    0.35     14   3.0 *******
    0.4      17   3.7 ********
    0.45     17   3.7 ********
    0.5       9   1.9 ****
    0.55     20   4.3 **********
    0.6       8   1.7 ****
    0.65     18   3.9 *********
    0.7      15   3.2 *******
    0.75     28   6.0 **************
    0.8      30   6.5 ***************
    0.85     45   9.7 ***********************
    0.9      24   5.2 ************
```

Keski-Suomi on vähän samantapainen, mutta ehdottoman varmoja luokituksia ei ole yhtään (maksimiarvo on alle yhden).

```
Variable: P3        Posterioritodennäköisyys (Pohjois)
min=0        in obs.#180 (Kökar)
max=1        in obs.#19 (Enontekiö)
mean=0.11422  stddev=0.223501 skewness=3.107988 kurtosis=8.916378
lower_Q=0.023115 median=0.046131 upper_Q=0.1045
up.limit      f     %      *=8 obs.  class width=0.05
    0.05    252  54.3 ******************************
    0.1      92  19.8 ***********
    0.15     50  10.8 ******
    0.2      17   3.7 **
    0.25     10   2.2 *
    0.3       3   0.6 :
    0.35      2   0.4 :
    0.4       0   0.0
    0.45      4   0.9 :
    0.5       4   0.9 :
    0.55      0   0.0
    0.6       4   0.9 :
    0.65      1   0.2 :
    0.7       1   0.2 :
    0.75      0   0.0
    0.8       1   0.2 :
    0.85      1   0.2 :
    0.9       0   0.0
    0.95      2   0.4 :
    1         9   1.9 *
    1.05     11   2.4 *
```

Äskeisiä alueita vierastanut Enontekiö löytää nyt paikkansa. Pohjois-Suomen jakauma on varsin erinäköinen kuin aiemmat: varmoja kuulumisia on vähän.

```
Variable: P4        Posterioritodennäköisyys (Saaristo)
min=0        in obs.#19 (Enontekiö)
max=0.999994 in obs.#180 (Kökar)
mean=0.03479  stddev=0.167177 skewness=5.104379 kurtosis=24.82912
up.limit      f     %      *=8 obs.  class width=0.05
    0        17   3.7 **
    0.05    423  91.2 **************************************************
    0.1       4   0.9 :
    0.15      1   0.2 :
    0.2       0   0.0
    0.25      1   0.2 :
    0.3       1   0.2 :
    0.35      1   0.2 :
    0.4       0   0.0
    0.45      0   0.0
    0.5       0   0.0
    0.55      1   0.2 :
    0.6       0   0.0
    0.65      0   0.0
    0.7       1   0.2 :
    0.75      0   0.0
    0.8       0   0.0
    0.85      4   0.9 :
    0.9       0   0.0
    0.95      0   0.0
    1        10   2.2 *
```

Saaristo-Suomen jakauma on vielä karumpi kuin Pohjois-Suomen.

Edellä olevien luokittelutaulukoiden perusteella Saaristo-Suomen 16 kunnasta 15 tulee luokitelluksi oikein mutta yksi väärin (Keski-Suomeen). Bayes-periaatteen (ryhmittäiset kovarianssit) taulukon mukaisesti yksi kunta vastaavasti sijoittuu Ruuhka-Suomesta Saaristo-Suomeen. Mikä on tämä vaihtokauppa? Selvitetään se aineistosta.

```
CASES=Alue:Saaristo IND=Bayes2,2 VARS=Kunta,P1,P2,P3,P4
FILE LOAD KUNNAT
 Kunta                    P1          P2          P3          P4
 Sund                  0.439       0.442       0.083       0.037
```

Sundin kunnan elinkeinorakenne poikkeaa ilmeisesti muista Ahvenanmaan kunnista. Se on luokiteltu miltei samalla todennäköisyydellä sekä Ruuhka- että Keski-Suomeen, mutta hädintuskin lainkaan Saaristo-Suomeen.

```
CASES=Alue:Ruuhka IND=Bayes2,4 VARS=Kunta,P1,P2,P3,P4
FILE LOAD KUNNAT
 Kunta                    P1          P2          P3          P4
 Kauniainen            0.197       0.000       0.000       0.802
```

Kauniainen puolestaan näyttäisi muistuttavan Ahvenanmaan kuntia jopa suuressa määrin. Luokittelun perusteella tulos olisi aivan selvä.

SAS:issa erotteluanalyysiin on useita proseduureja, mm. CANDISC ja DISCRIM.

Ryhmittelymenetelmät

Erotteluanalyysin yhteydessä kysyttiin, mikä erottaa tunnetut ryhmät toisistaan. Jos ryhmiä ei tunnetakaan ennalta mutta epäillään aineiston olevan jollain tavoin heterogeeninen, ollaan paljon hankalamman tilanteen edessä.

Ryhmittelyanalyysi *(cluster analysis)* on yhteisnimitys monentyyppisille menetelmille, joilla aineiston heterogeenisuutta pyritään tuomaan esille. Osa menetelmistä on luonteeltaan hierarkisia. Ne soveltuvat parhaiten verraten pienille aineistoille, joissa havainnot on luontevaa yksilöidä. Suuremmille aineistoille on omat hieman toisentyyppiset menetelmänsä.

Pohditaan erilaisia ryhmittelykeinoja pienellä esimerkkiaineistolla, jonka on alunperin esittänyt *Seppo Mustonen* monimuuttujamenetelmien kursseillaan. Helsingin Sanomat kokosi keväällä 1996 kyselykaavakkeella EU-maiden ja EU-komission käsityksiä EU:n tulevaisuuden tavoitteista. Maiden Brysselin edustustot ottivat kantaa 15 kysymykseen 5-portaisella asteikolla.

```
FILE STATUS EUMAAT
EU-maitten kannat keväällä 1996
5=Voimakas kannatus, 4=Kannattaa, 3=Kiinnostunut, 2=Ei hyväksy, 1=Vastustaa
FIELDS: (active)
   1 SA_  12 Maa
   2 NA_   1 TiivisEU Pyrittävä yhä tiiviimpään integraatioon (#)
   3 NA_   1 ItäEU    Laajennuttava itä- ja keski-Eurooppaan (#)
   4 NA_   1 Äänet    Äänimääriä muutettava ministerineuvostossa (#)
   5 NA_   1 Puhjoht  Puheenjohtajuusjärjestelmää syytä muuttaa (#)
   6 NA_   1 Komissio Komission kokoa ratkaisevasti rajoitettava (#)
   7 NA_   1 Parlvalt Parlamentille lisää valtaa (#)
   8 NA_   1 Yhtpäät  Yhteispäätösmenettelyä lisättävä (#)
   9 NA_   1 Määräenm Määräenemmistöpäätösten osuutta lisättävä (#)
  10 NA_   1 Työttöm  Työttömyyden torjunta perussopimukseen (#)
  11 NA_   1 Päätulko Määräenemmistöpäätöksiä lisättävä ulkopolitiikassa (#)
  12 NA_   1 Ulkomin  Ulkoministeriä tarvitaan (#)
  13 NA_   1 EU\WEU   EU ja WEU yhdistettävä (#)
  14 NA_   1 Siirtol  Siirtolaisuusasiat unionin toimivaltaan (#)
  15 NA_   1 Poliisi  Poliisiasiat unionin toimivaltaan (#)
  16 NA_   1 Schengen Schengenin sopimusta laajennettava (#)
END
Survo data file EUMAAT: record=53 bytes, M1=24 L=64  M=16 N=14
```

Aineisto on niin pieni, että se voidaan ottaa kokonaan näkyviin. (Paljoa tämän enempää ei HS ollut aineistosta onnistunutkaan saamaan esille.) Syvemmälle kaivautuminen edellyttää usean muuttujan yhtaikaisia tarkasteluja.

```
FILE LOAD EUMAAT CUR+1
 Maa          T I Ä P K P Y M T P U E S P S
 Saksa        5 5 4 4 5 5 5 5 2 5 3 5 5 3 5
 Ranska       5 5 5 3 5 5 5 5 1 5 5 4 4 2 4
 Britannia    1 5 4 3 4 1 2 1 1 1 1 1 1 1 2
 Italia       5 4 5 3 4 4 4 4 2 4 3 4 4 3 3
 Espanja      4 3 5 3 2 2 3 3 4 4 3 4 2 2 2
 Suomi        2 5 1 1 1 3 3 4 5 3 1 2 2 2 4
 Ruotsi       3 5 1 2 1 2 3 3 5 3 3 2 4 2 4
 Itävalta     5 5 2 3 3 4 4 4 5 4 3 2 4 3 4
 Tanska       2 5 3 3 5 3 3 4 5 2 3 3 2 2 3
 Irlanti      5 5 3 1 1 2 3 5 3 2 3 1 4 4 4
 Benelux      5 5 3 4 1 4 5 4 4 5 2 5 5 3 5
 Portugali    4 3 3 3 1 2 3 4 4 4 2 4 4 4 4
 Kreikka      5 5 2 2 2 5 5 4 5 3 2 4 4 2 3
 Komissio     5 5 4 2 2 4 5 4 5 5 1 5 4 2 5
```

Visuaalinen ryhmittely

Faktori- ja pääkomponenttianalyysejakin voidaan ajatella ryhmittelymenetelminä, sillä niissä ryhmitellään muuttujia. Sisällöllisesti loogisempi valinta tässä esimerkissä olisi jälleen faktorianalyysi. Nyt muuttujia (väitteitä) on enemmän kuin havaintoja (edustustoja), joten faktorianalyysi ei onnistu ainakaan suurimman uskottavuuden menetelmällä. Pääakselimenetelmällä ratkaisun saa aikaan, mutta näin pientä aineistoa voidaan visualisoida suoraankin useammalla vaihtoehtoisella tavalla.

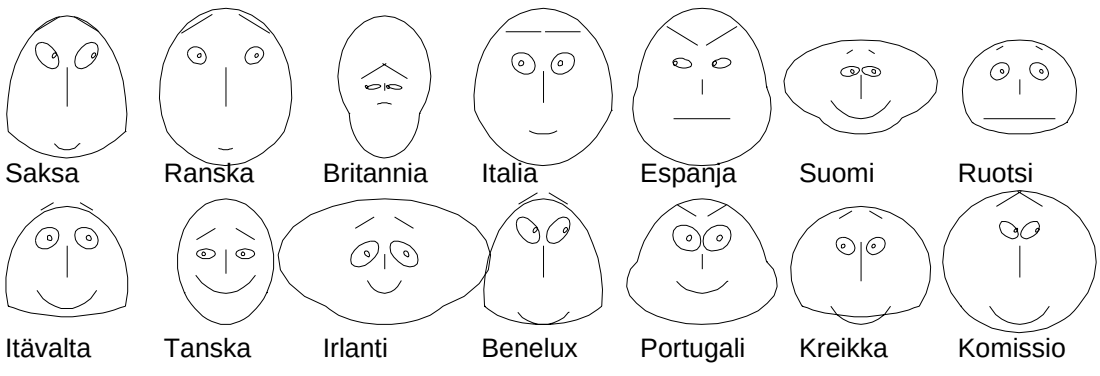
Herman Chernoffin kehittämä moniulotteinen kuvaustekniikka perustuu siihen, että ihminen kykenee tunnistamaan muita ihmisiä hyvin nopeasti kasvonpiirteiden avulla. Aineiston havainnot voidaan kuvata naamoina kytkemällä muuttujia eri kasvonpiirteisiin. Kuvan perusteella saadaan läpileikkaus koko aineistosta, ja voidaan ryhmitellä havaintoja niiden visuaalisten ominaisuuksien mukaan.

Chernoffin alkuperäisessä mallissa (josta on kehitetty lukuisia muunnelmia) on mahdollisuus 18 kasvonpiirteen käyttöön. EU-aineistossa on vain 15 muuttujaa, joten saadakseni kaikki piirteet varioimaan olen laittanut kolme ensimmäistä muuttujaa mukaan toiseen kertaan listan loppuun. Luettelo piirteistä saadaan automaattisesti, eikä siihen tarvitse kuin lisäillä muuttujien nimiä.

```
GPLOT EUMAAT / TYPE=FACES LABEL=Maa
```

```
VARIABLES: xmin      xmax     Features                        fmin fmax
TiivisEU  1*        5**      Radius_to_corner_of_face_OP      0.6  1.0
ItäEU     3*        5**      Angle_of_OP_to_horizontal        0.0  0.6
Äänet     1*        5**      Vertical_size_of_face_OU         0.6  1.0
Puhjoht   1*        4**      Eccentricity_of_upper_face       0.5  1.5
Komissio  1*        5**      Eccentricity_of_lower_face       0.5  1.5
Parlvalt  1*        5**      Length_of_nose                   0.1  0.5
Yhtpäät   2*        5**      Vertical_position_of_mouth       0.2  0.8
Määräenm  1*        5**      Curvature_of_mouth_1/R          -4.0  4.0
Työttöm   1*        5**      Width_of_mouth                   0.2  1.0
Päätulko  1*        5**      Vertical_position_of_eyes        0.0  0.4
Ulkomin   1*        5**      Separation_of_eyes               0.3  0.8
EU\WEU    1*        5**      Slant_of_eyes                   -0.5  0.5
Siirtol   1*        5**      Eccentricity_of_eyes             0.3  1.0
Poliisi   1*        4**      Size_of_eyes                     0.1  0.2
Schengen  2*        5**      Position_of_pupils              -0.1  0.1
TiivisEU  1*        5**      Vertical_position_of_eyebrows    0.2  0.4
ItäEU     3*        5**      Slant_of_eyebrows               -0.5  0.5
Äänet     1*        5**      Size_of_eyebrows                 0.1  0.5
END
```

Chernoff's faces: EUMAAT



Kuvan ja aineiston tarkastelun avulla näkee mitä moninaisimpia asioita: Britannia näyttää todella epäluuloiselta, eikä syyttä: se vastustaa lähes kaikkia asioita. Suomella ja Irlannilla on paljon yhteistä, ne mm. vastustavat puheenjohtajuusjärjestelmän muutoksia. Ruotsin neutraalius monessa asiassa näkyy suorastaan naamasta. Saksan ja Benelux-maiden edustustot ovat keskenään samoilla linjoilla lähes kaikista asioista.

Chernoffin naamoja on käytetty menestyksekkäästi mm. konkurssikypsien yritysten seulontaan tilinpäätöstietojen perusteella. Itse olen havainnollistanut naamojen avulla mm. eri asteisia ongelmaluokituksia ja niihin mahdollisesti vaikuttavia tekijöitä erään lastenpsykiatrisen tutkimuksen yhteydessä. Havaintomäärältään pienten aineistojen kohdalla naamakuvat tarjoavat vähintäänkin varteenotettavan keinon tutustua aineistoon.

Naamakuvien lisäksi on olemassa joukko muitakin moniulotteisen aineiston kuvaustapoja, mm. profiili- ja tähtikuvat sekä Andrewsin käyrät. Naamat lienevät näistä ainakin ilmeikkäimpiä.

Hierarkinen ryhmittely

Toinen tapa nähdä havaintojen välisiä yhteyksiä on soveltaa hierarkista ryhmittelyä. Sen päätuloksena piirretään *dendrogrammiksi* kutsuttu puumainen kuva, joka ainakin biologeille on ennestään tuttu taksonomian yhteydestä.

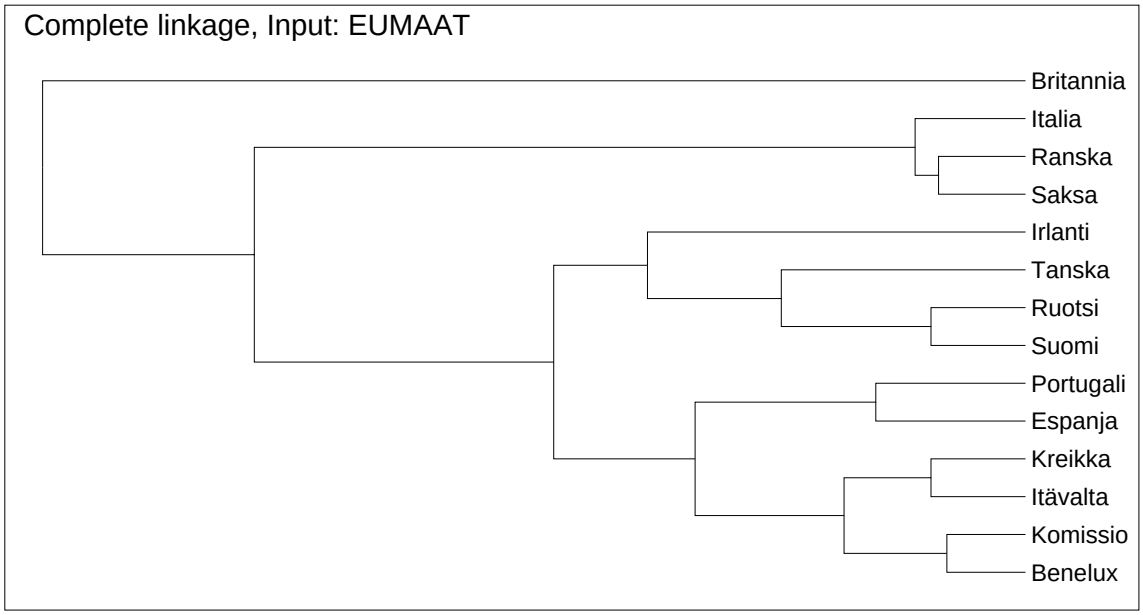
Hierarkiset ryhmittelyt ovat luonteeltaan heuristisia, siis juuri mitään tilastollisia kriteerejä ei ole. Kyseessä on teema ja muunnelmat, sillä ensin pitää päättää miten havaintojen välisiä etäisyyksiä *(distance)*, läheisyyksiä *(proximity)*, erilaisuuksia *(dissimilarity)* tai samankaltaisuuksia *(similarity)* mitataan. On olemassa runsaasti erilaisia etäisyysmittoja erikseen jatkuville ja diskreeteille muuttujille. Lopuksi on vielä valittava sopiva ryhmittelyalgoritmi. Näiden yhdistelminä saadaan laaja joukko vaihtoehtoja, joista mikään ei periaatteessa ole oikeampi kuin toinen. Hierarkinen ryhmittely edellyttääkin tyypillisesti useiden eri vaihtoehtojen kokeilemista.

Eri aloilla saattaa olla vakiintuneita käytäntöjä siihen mitä etäisyysmittoja käytetään missäkin tilanteissa. Tässä esimerkissä voidaan valita yhtä hyvin mitä tahansa vaihtoehtoja. Ryhmittelyalgoritmeista tavallisin on ns. yksinkertainen ketjutus *(single linkage)*, joka tunnetaan myös lähimmän naapurin *(nearest neighbour)* menetelmän nimellä. Sitä ei useinkaan kannata soveltaa, koska havainnot ketjuuntuvat liiaksi toisiinsa ja ryhmiä on siten vaikea erottaa. Parempia ovat esim. täydellinen ketjutus *(complete linkage)* ja minimivarianssikriteeriin perustuva Wardin menetelmä.

Ryhmitellään nyt EU-maat käyttäen etäisyysmittana tavallisen euklidisen etäisyyden neliötä (erot hieman selvemmin esille) ja ryhmittelyalgoritmina täydellistä ketjutusta.

```
HCLUSTER EUMAAT,CUR+2 / METHOD=COMPLETE_LINKAGE
```

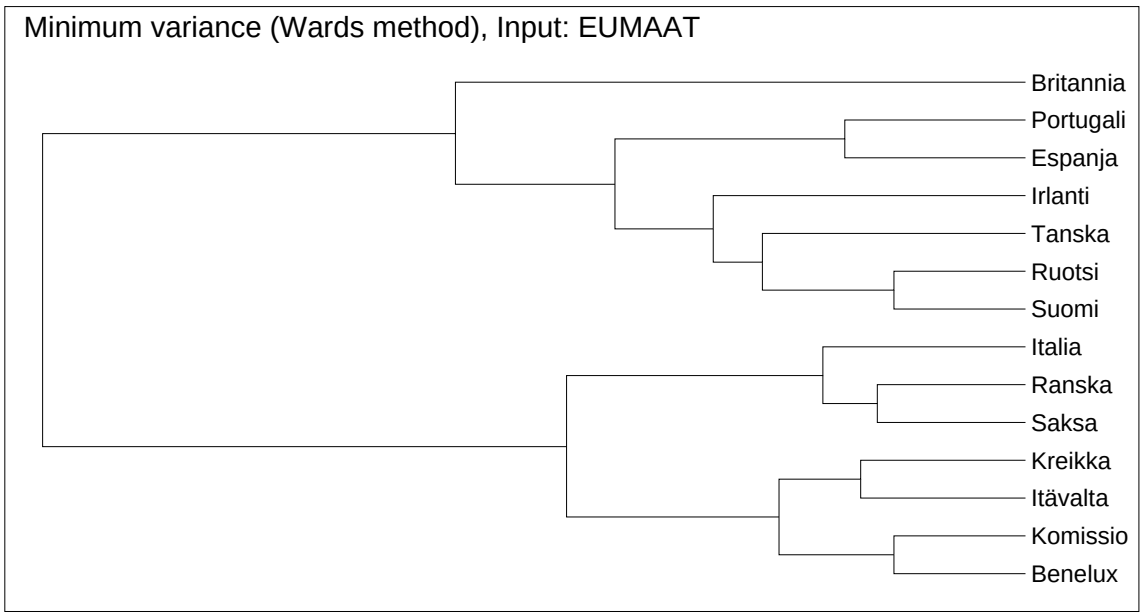
Analyysin oleellisin tulos on kuva:



Hierarkista ryhmittelyä voidaan tehdä joko jakavilla *(divisive)* tai kasaavilla *(agglomerative)* algoritmeilla. Edellisessä kaikki havainnot ovat aluksi samassa ryhmässä (kuvan vasen reuna), ja lopuksi jokainen muodostaa oman ryhmänsä (kuvan oikea reuna). Kasaava algoritmi etenee vain toisin päin. Lopputuloksena saatavasta dendrogrammista nähdään joka tapauksessa mitkä havainnot ovat lähimpänä toisiaan, ja voidaan arvioida oikeaa ryhmien lukumäärää, mikäli siitä ollaan kiinnostuneita.

Britannian erimielisyys kuvastuu tästäkin. Sen etäisyys muihin on suurin. Jos sovelletaan ns. *city block* -etäisyyttä ja Wardin menetelmää, päädytään hieman toisennäköiseen dendrogrammiin.

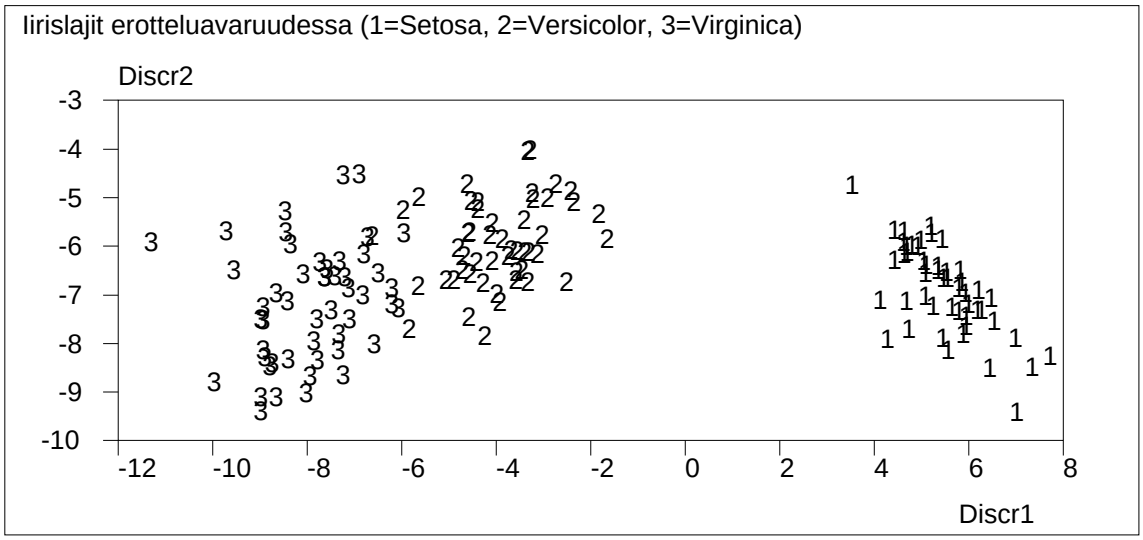
```
HCLUSTER EUMAAT,END+2 / METHOD=MINIMUM_VARIANCE DISTANCE=CITY_BLOCK
```



Tämän kuvan perusteella EU-edustustot voisi jakaa kahteen ryhmään.

Muita ryhmittelymenetelmiä

Tutkitaan seuraavaksi kuuluisaa aineistoa, jota *R. A. Fisher* aikoinaan käytti kehittämänsä erotteluanalyysin havainnollistamiseen. Aineistossa on tietoja kolmen iirislajin terälehdistä tehdyistä mittauksista, neljä muuttujaa ja 50 havaintoa kustakin lajista. Erotteluanalyysin perusteella ne eroavat toisistaan selvästi, varsinkin yksi lajeista poikkeaa täysin kahdesta muusta. Tämä näkyy kuvasta, jossa on piirretty erottelumuuttujat vastakkain. Havaintopisteinä ovat lajien tunnukset (1=Setosa, 2=Versicolor ja 3=Virginica).



Todellisessa tilanteessa ei yleensä ole mitään oikeaa luokitusta johon saatua ryhmitystä voitaisiin verrata, koska tällöin voitaisiin yhtä hyvin käyttää tunnettua luokitusta ja soveltaa erotteluanalyysia. Tässä yhteydessä on kuitenkin hyödyllistä katsoa, miten erilaisilla ryhmittelymenetelmillä saadut tulokset vastaavat todellisuutta.

Käytännössä suosituin ryhmittelymenetelmä tunnetaan nimellä *k-means*. Nimensä mukaisesti siinä toimitaan ryhmien keskiarvopisteiden kanssa. Menetelmän tunnettu heikkous on se, että ryhmittelytulokset riippuvat alkuryhmittelystä. Niinpä menetelmää onkin tapana käyttää siten että toistetaan sitä erilaisilla alkutilanteilla ja valitaan paras saaduista ratkaisuista.

Uudempi medoidiryhmittely on tässä suhteessa parempi, sillä se antaa yksikäsitteisiä tuloksia ja on muutenkin robustimpi kuin *k-means*. Tämä perustuu siihen, että keskiarvojen sijasta siinä toimitaan mediaanien pohjalta.

Ryhmittelyn pohjaksi tarvitaan havaintojen välinen etäisyysmatriisi, joten jälleen tulokset ovat riippuvaisia etäisyysmitan valinnasta. Kokeillaan aluksi tavallisia euklidisia etäisyyksiä.

```
MASK=AAAA------
DIST IRIS,D
```

Etäisyysmatriisi D on siis dimensioiltaan 150x150. Kovin valtavia aineistoja ei voida tätä kautta tutkia, sillä etäisyysmatriisin on mahduttava kerralla keskusmuistiin. Nykyisillä koneilla voidaan kuitenkin helposti analysoida muutaman tuhannen havainnon välisiä (siis miljoonien alkioiden) etäisyysmatriiseja.

Vilkaistaan matriisin oikeaa alanurkkaa, josta näkyy sen etäisyysmatriisille tyypilliset ominaisuudet, symmetrisyys ja "onttous". Jälkimmäinen tarkoittaa sitä että matriisin diagonaalilla on nollaa (havainnon etäisyys itseensä).

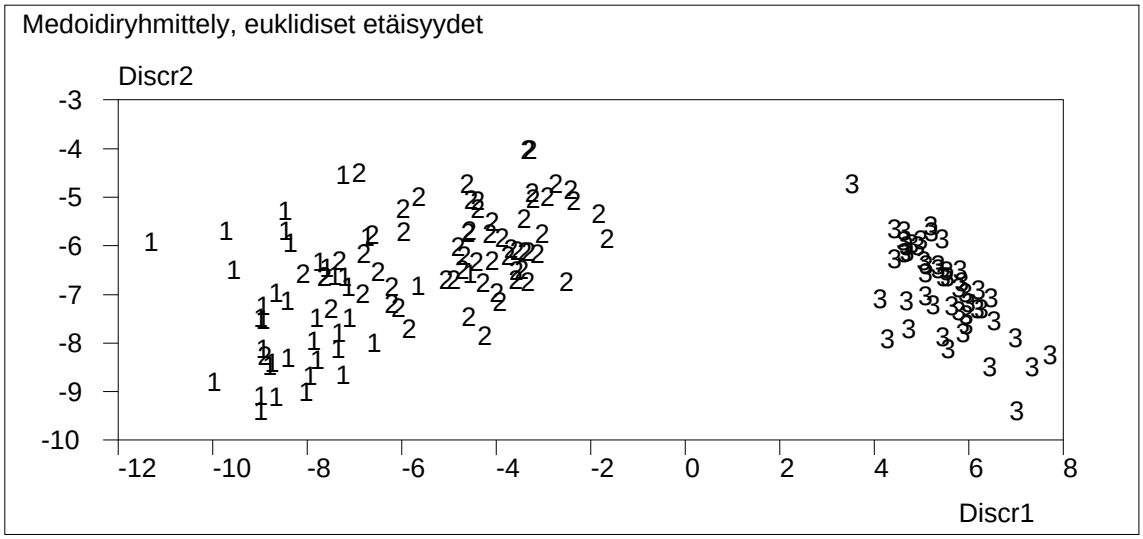
```
MAT LOAD D(146:150,142:150),##.##,CUR+1
MATRIX D
Euclidean
///        142   143   144   145   146   147   148   149   150
146      43.97 11.83  6.78  5.57  0.00 45.93 44.61 41.64  5.57
147       9.33 35.24 49.42 44.98 45.93  0.00  8.83  7.87 43.74
148       3.32 33.67 47.10 42.63 44.61  8.83  0.00  5.48 41.73
149       3.61 30.36 44.32 40.06 41.64  7.87  5.48  0.00 38.72
150      40.84  8.66  6.08  4.24  5.57 43.74 41.73 38.72  0.00
```

Korrelaatiomatriisissahan on vastaavasti diagonaalilla ykköset (muuttujan korrelaatio itsensä kanssa), eli sitä voidaan kutsua samankaltaisuusmatriisiksi. Ryhmittely- ja muissa menetelmissä lähtökohtana on lähes aina etäisyys- tai erilaisuusmatriisi. Näitä voi muunnella eri suuntiin monotonisilla muunnoksilla, yksinkertaisimmillaan vähentämällä arvot ykkösestä.

Tehdään ryhmittelyanalyysi medoidimenetelmällä edellä lasketusta etäisyysmatriisista. Ryhmien lukumäärää ei voi välttämättä ennalta tietää, joten on vain kokeiltava. Ryhmittelyä auttaa suuresti, jos ensin piirtää kuvan. Tässä on kuvat piirretty erottelumuuttujien avulla, mutta todellisessa tilanteessa voi käyttää vaikkapa paria ensimmäistä pääkomponenttia. Myös itse ryhmittelyn voi tehdä esim. pääkomponentti- tai faktoripistemäärien perusteella. Tämä on sitä järkevämpää mitä useampaan muuttujaan ryhmittely aiotaan perustaa.

```
MASK=--------GS
DCLUSTER IRIS,D,CUR+2 / GROUPS=3

Cluster analysis by medoids of Kaufman and Rousseeuw (1987)
Data IRIS  N=150
Group Medoid       n Mean (of silhouette values)
 1    34          38 0.451
 2    56          62 0.417
 3    137         50 0.798
Mean of all silhouette values is 0.553
```

Menetelmä on muodostanut kolme ryhmää, joissa on 38, 62 ja 50 havaintoa. Arvatenkin kolmas tarkoittaa Setosa-lajia, sillä se on niin kaukana muista että käytännössä mikä tahansa ryhmittelymenetelmä löytää sen oikein. Muiden osalta jo havaintomäärät paljastavat ettei oikeita ryhmiä löydetä. Havaintokohtaisten silhuettiarvojen keskiarvot kertovat että ryhmittely on ollut vaikeinta ryhmien 1 ja 2 osalta. Kuvasta näkyy että vasemmanpuoleisen havaintojoukon jako kahteen osaan ei ole aivan helppoa.



Ryhmien numerointi ei tietenkään välttämättä vastaa edellä käytettyä vaan menetelmä numeroi ryhmät mielivaltaisesti. Ne voisi muuttaa vastaamaan oikeita mutta en ole sitä tässä katsonut tarpeelliseksi tehdä. Olennaista on se, miten hyvin ryhmät ylipäätään löytyvät.

Jos valitaan etäisyysmitaksi korrelaatiot (ykkösestä vähennettyinä) ja tehdään ryhmittely uudelleen, saadaan huomattavasti parempi lopputulos.

```
MASK=AAAA----
DIST IRIS,D / MEASURE=Correlation
```

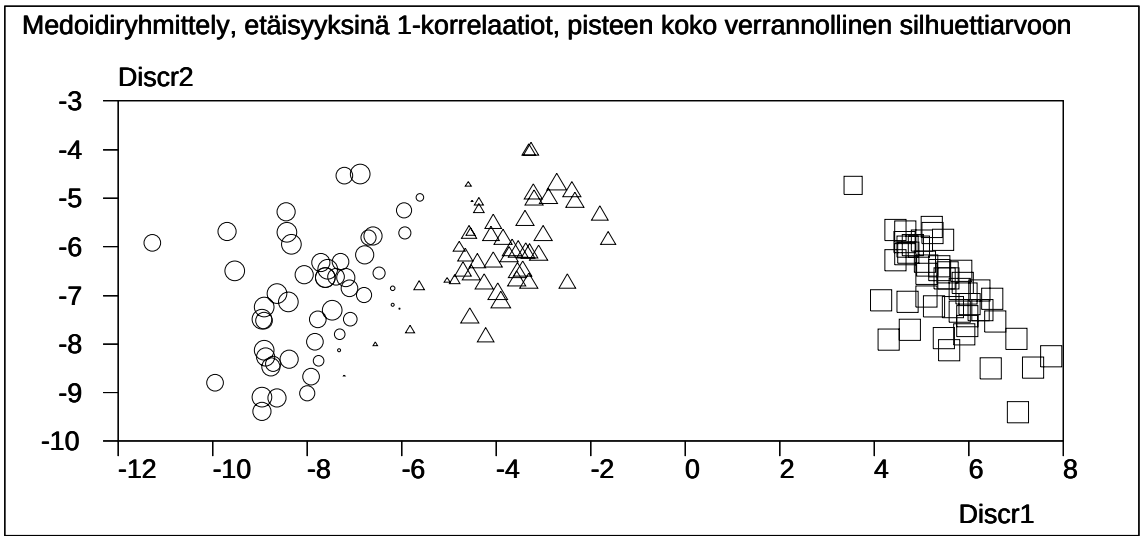
Vilkaistaan taas etäisyysmatriisin oikeaa alanurkkaa.

```
MAT LOAD D(146:150,142:150),##.##,CUR+1
MATRIX D
Correlation
///        142   143   144   145   146   147   148   149   150
146       0.40  0.03  0.00  0.01  0.00  0.34  0.44  0.33  0.01
147       0.00  0.17  0.28  0.25  0.34  0.00  0.01  0.00  0.24
148       0.00  0.25  0.37  0.34  0.44  0.01  0.00  0.01  0.32
149       0.00  0.16  0.28  0.25  0.33  0.00  0.01  0.00  0.23
150       0.29  0.01  0.00  0.00  0.01  0.24  0.32  0.23  0.00
```

Tehdään ryhmittely samaan tapaan kuin edellä.

```
MASK=--------GS
DCLUSTER IRIS,D,CUR+2 / GROUPS=3

Cluster analysis by medoids of Kaufman and Rousseeuw (1987)
Data IRIS  N=150
Group Medoid       n Mean (of silhouette values)
 1    82          50 0.747
 2    42          50 0.693
 3    124         50 0.977
Mean of all silhouette values is 0.806
```

Nyt löytyy kolme oikeankokoista ryhmää, ja silhuettiarvot ovat selvästi parempia. Kuvassa havaintopisteet on kuvattu ryhmittäin eri merkeillä. Mitä isompi merkki, sitä suurempi silhuettiarvo. Pienimpien pisteiden ryhmittelyyn liittyy eniten epävarmuutta.



Ristiintaulukoimalla ryhmittelyn ja alkuperäiset tiedot nähdään että molemmista vasemmanpuoleisen havaintojoukon ryhmistä on siirtynyt toiseen ryhmään kolme havaintoa.

```
TAB IRIS END+2 / VARIABLES=Group,specname CHI2=- LABELS=0
    Group=1,1,2,3 specname=/Setosa,/Versicolor,/Virginica

            Group   1   2   3
specname    *****
/Setosa             0   0  50
/Versicolor         3  47   0
/Virginica         47   3   0
```

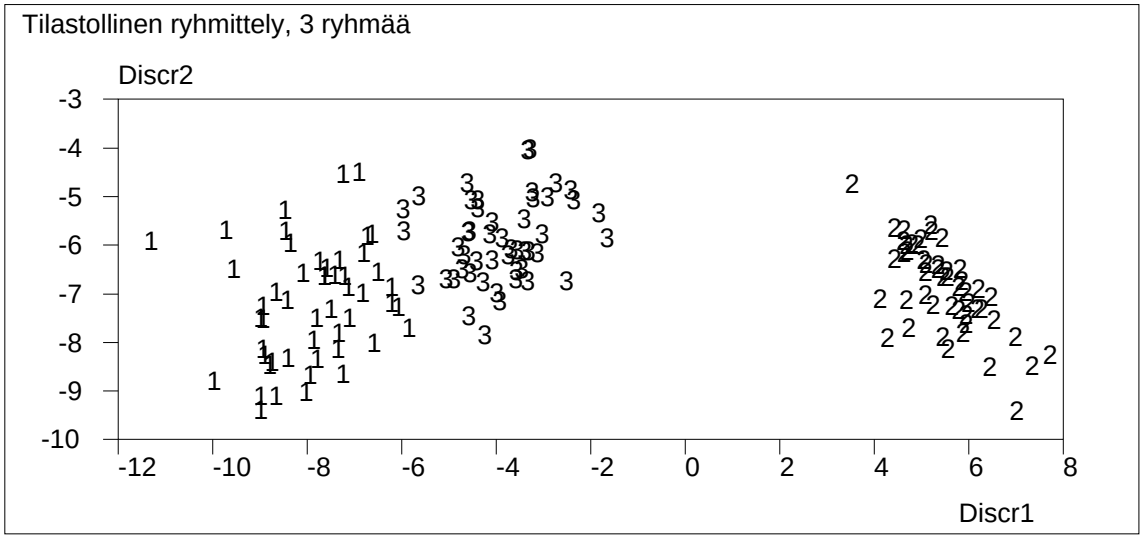
Viimeiseksi tässä esiteltävä ryhmittelymenetelmä on siitä harvinainen, että sillä on tilastollinen perusta. Niinpä sitä voidaankin kutsua tilastolliseksi ryhmittelyanalyysiksi. Edellä mainitun *k-means*-ryhmittelyn tapaan havainnot jaetaan aluksi ryhmiin umpimähkäisesti, minkä jälkeen niitä siirrellään systemaattisesti ryhmistä toisiin. Tavoitteena on luoda mahdollisimman homogeenisia ryhmiä. Lopputulos saavutetaan, kun yhdenkään havainnon siirto ei enää paranna tilannetta. Paras ratkaisu ei ole taattu, sillä satunnainen alkuryhmitys vaikuttaa lopputulokseen. Menettely onkin syytä toistaa useita kertoja erilaisilla aloituksilla ja valita saaduista tuloksista paras.

Faktori- ja erotteluanalyysin tapaan tilastollisessa ryhmittelyanalyysissa on taustalla oletus havaintojen multinormaalijakaumasta. Oletuksen pätiessä paras mitta ryhmien homogeenisuudelle on ns. *Wilksin lambda*, jota sovelletaan yleisesti myös erotteluanalyysin yhteydessä.

Tässäkin on todellisuudessa kokeiltava eri ryhmälukuja. Katsotaan miltä näyttää kolmella ryhmällä, jotta voidaan vertailla. Nyt ei tarvitse laskea etäisyysmatriisia, joten tilastollista ryhmittelyä voi harrastaa kuinka suurilla aineistoilla tahansa. Laitetaan ohjelma toistamaan ryhmittelyä peräti 1000 kertaa.

```
MASK=AAAA----G-
CLUSTER IRIS,CUR+2 / GROUPS=3 TRIALS=1000

Stepwise cluster analysis by Wilks' Lambda criterion
Data IRIS  N=150
Variables: sepallen, sepalwid, petallen, petalwid
Best clusterings found in 1000 trials are saved as follows:
 Lambda          freq  Grouping var
0.0001680966    120   Group
```

Parhaaseen eli pienimmän lambda-arvon ryhmitykseen päädyttiin 120 kertaa. Kuvan perusteella ratkaisu näyttää varsin hyvältä, ryhmien numerointi vain sattuu taas olemaan eri kuin edellä.



Taulukoimalla selviää että vain kolme havaintoa on sijoitettu väärin.

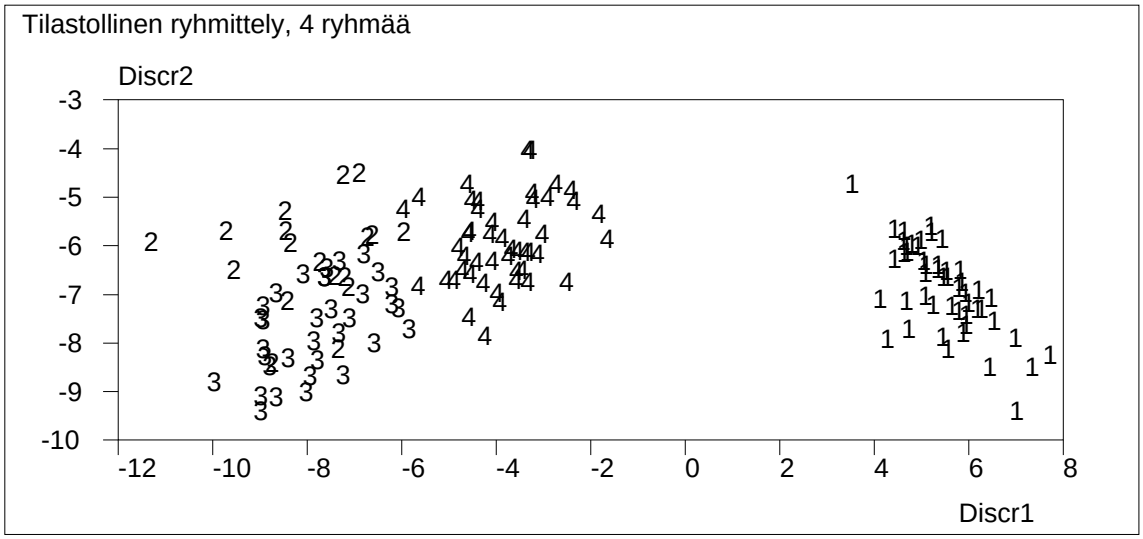
	Group	1	2	3
specname	*****			
/Setosa		0	50	0
/Versicolor		**2**	0	48
/Virginica		49	0	**1**

Tilastollinen ryhmittelyanalyysi toimii siis tässä aineistossa erinomaisesti, varsinkin kun oikea ryhmien lukumäärä tiedettiin ennalta. Yleisesti ei voi koskaan tietää mikä ryhmittely toimii missäkin tilanteessa mitenkin hyvin ja mikä on oikea ryhmälukumäärä. Useimmiten on vain kokeiltava monenlaisia lähestymistapoja ja valittava se lopputulos mikä vaikuttaa parhaalta.

Katsotaan vielä lopuksi miten käy jos koetetaankin löytää neljä ryhmää.

```
MASK=AAAA----G-
CLUSTER IRIS,CUR+1 / GROUPS=4 TRIALS=1000
Stepwise cluster analysis by Wilks' Lambda criterion
Data IRIS  N=150
Variables: sepallen, sepalwid, petallen, petalwid
Best clusterings found in 1000 trials are saved as follows:
 Lambda          freq  Grouping var
0.0000701363    816   Group
```

Nyt 1000 kokeilun kuluessa on menetelmä päätynyt samaan lambda-arvoon yli 800 kertaa. Havaintojoukon reunalle on syntynyt uusi ryhmä. Tätä voisi todellisessa tilanteessa pitää aivan yhtä hyvänä ratkaisuna kuin aiempia.



Ryhmittelytulosten järkevyyttä kannattaa testata erotteluanalyysilla sekä taulukoimalla ryhmiä ristiin esimerkiksi taustamuuttujien ym. kanssa.

SAS:issa hierarkista ryhmittelyä tehdään CLUSTER-proseduurilla ja *k-means*-ryhmittelyä FASTCLUS-proseduurilla.

Kanoninen analyysi

Kanonista analyysia *(canonical analysis)* voidaan pitää regressioanalyysin yleistyksenä tilanteeseen jossa selitettäviä muuttujia on useita. Niistä kuten myös selittäjistä muodostetaan molemmista erikseen sellaisia lineaarikombinaatioita joiden välinen korrelaatio on mahdollisimman suuri. Menetelmää kutsutaan myös nimellä kanoniset korrelaatiot *(canonical correlations)*.

Esimerkkinä tarkastellaan Biometria-kirjassa esitettyä eteläsuomalaisen kuusen kasvuun, kokoon ja kuntoon liittyvää aineistoa. Erään koemetsikön puista on valittu satunnaisotannalla 30 puuta, joista on koko- ja kasvutunnuksina mitattu puun ikä (vuosia), pituus (m), viiden viimeisen vuoden pituuskasvu (m) ja tilavuuskasvu (dm^3). Elävien kuusien kuntoa on mitattu arvioimalla neulaskato, nilakerroksen vastus tai sähkönjohtokyky (kOhm) sekä neulasvuosikertojen määrä. Muuttujat on pyritty normalisoimaan erilaisilla muunnoksilla. Lisäksi ne on standardoitu (keskiarvot nollia, hajonnat ykkösiä).

```
FILE STATUS KUUSET
Kuusen kasvuun liittyvä aineisto / Biometria 3.painos s.493
FIELDS: (active)
   1 NA_   4 X1       Ikä
   2 NA_   4 X2       Pituus
   3 NA_   4 X3       Pituuskasvu
   4 NA_   4 X4       Tilavuuskasvu
   5 NA_   4 Y1       Neulaskato
   6 NA_   4 Y2       Sähkönjohtokyky
   7 NA_   4 Y3       Neulasvuosikerta
END
Survo data file KUUSET: record=55 bytes, M1=12 L=64  M=7 N=30
```

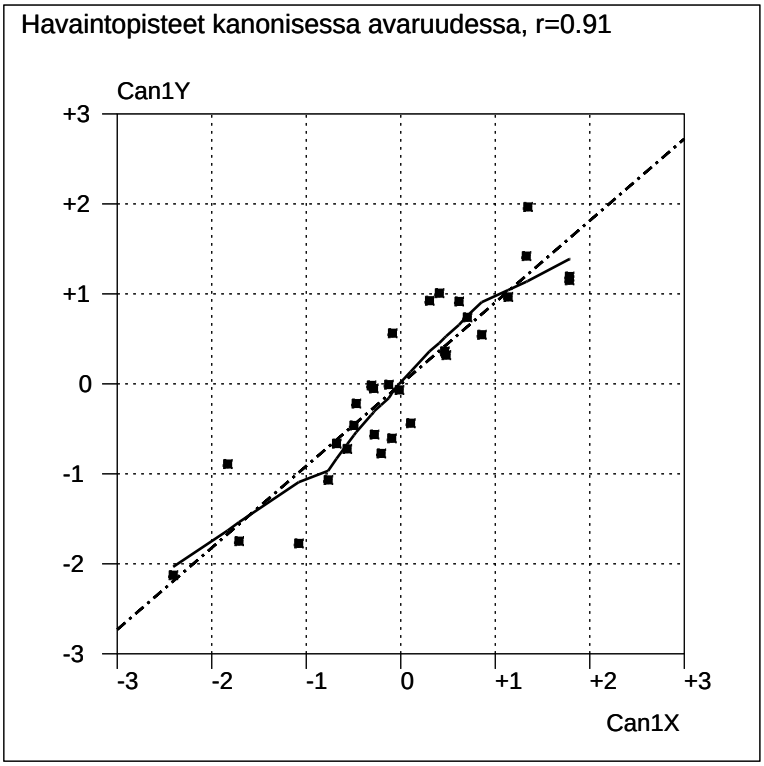
Tehdään kanoninen analyysi, jossa tutkitaan koko- ja kasvutunnusten yhteyksiä kuusten kuntoon. Selitettävien ja selittäjien roolit riippuvat tutkimusasetelmasta ja edellyttävät tietoa asian sisällöstä. Kuten yleensäkin, korrelaatiot kuvaavat vain yhteyksiä lineaarisen riippuvuuden mielessä. Kausaliteetin tulkinta jää tutkijalle. Kuntoon liittyvät muuttujat on merkitty Y:llä ikään kuin selitettäviksi, mutta kirjaimilla ei tässä yhteydessä ole niin paljoa väliä kuin regressioanalyysin puolella. Yhteyksiä voi tarkastella kummin päin tahansa.

```
MASK=XXXXYYY
CANON KUUSET CUR+2

Canonical analysis on KUUSET:
Correlation    CHI^2     P       df      (LCAN.M)
 1   0.9088   49.098  1.00000    12
 2   0.3629   5.4011  0.50651     6
 3   0.2686   1.8717  0.60775     2
Coefficients (LINCO) for canonical variables saved in XCOEFF.M,YCOEFF.M

Correlations of canonical variables with X variables XCAN.M
          CAN1   CAN2   CAN3
X1       -0.346 -0.748 -0.063  Ikä
X2       -0.770 -0.134 -0.622  Pituus
X3       -0.987  0.101  0.104  Pituuskasvu
X4       -0.794 -0.359 -0.278  Tilavuuskasvu

Correlations of canonical variables with Y variables YCAN.M
          CAN1   CAN2   CAN3
Y1        0.962 -0.253  0.106  Neulaskato
Y2        0.806  0.536  0.250  Sähkönjohtokyky
Y3       -0.555 -0.467  0.688  Neulasvuosikerta
```

Vain ensimmäisen kanonisen muuttujaparin välinen korrelaatio (0.91) on merkitsevä. Jälkimmäisistä nähdään kuitenkin jotain suuntaa antavaa, esimerkiksi puun ikä tulisi enemmän esille toisella kanonisella X-muuttujalla. Se ei kuitenkaan ole ainakaan tällä aineistolla kiinnostavaa, koska vastaava kanoninen korrelaatio on vain 0.36.

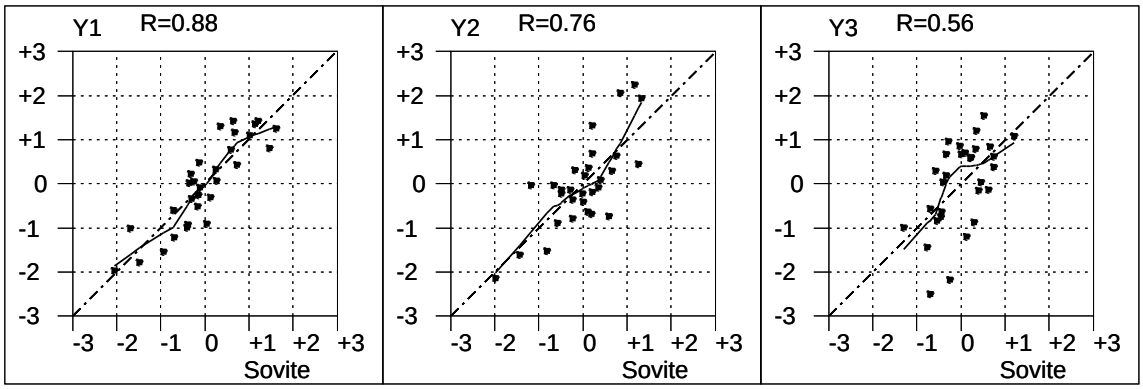
Parhaalla kanonisella X-muuttujalla korostuu erityisesti pituuskasvu, mutta myös tilavuuskasvu ja pituus, kaikki negatiivisina korrelaatioina. Vastaavalla Y-muuttujalla korostuvat muut paitsi neulasvuosikerta. Huonot kasvuluvut näyttävät siis olevan yhteydessä neulaskatoon ja puun huonosta kunnosta indikoivaan suureen sähkönjohtokykyyn.

Lasketaan havaintokohtaiset kanoniset pistemäärät painotettuina summina alkuperäisistä muuttujista ja piirretään kanoniset pistemäärät vastakkain. Nyt pistemäärät eivät olekaan korreloimattomia kuten faktori- ja pääkomponenttianalyyseissa vaan ne korreloivat juuri vastaavan kanonisen korrelaation verran. Lisätään kuvaan regressiosuora ja -tasoitus.



Vastaavat tarkastelut tavallisen regressioanalyysin avulla ovat kömpelöitä, koska selitettävä ilmiö on moniulotteinen. Tehdään vertailun vuoksi kolme regressioanalyysia (kullekin Y-muuttujalle erikseen) ja piirretään kustakin analyysista X-muuttujien lineaarikombinaatio (sovite) vastakkain selitettävän muuttujan kanssa.

```
MASK=XXXXY----P
REGDIAG KUUSET CUR+2

Regression diagnostics on data KUUSET: N=30
Regressand Y1        # of regressors=5 (Constant term included)
Condition number of scaled X: k=4.37129
Variable Regr.coeff. Std.dev.     t
Constant  0.0000589  0.0936801     0.0006
X1       -0.0164103  0.1033751    -0.1587  Ikä
X2       -0.2487048  0.1870388    -1.3297  Pituus
X3       -0.8097522  0.1453284    -5.5719  Pituuskasvu
X4        0.1538139  0.2011566     0.7646  Tilavuuskasvu
Variance of regressand Y1=1.000062626 df=29
Residual variance=0.263278628 df=25
R=0.8792 R^2=0.7730 Durbin-Watson=2.246

MASK=XXXX-Y---P
REGDIAG KUUSET CUR+2

Regression diagnostics on data KUUSET: N=30
Regressand Y2        # of regressors=5 (Constant term included)
Condition number of scaled X: k=4.37129
Variable Regr.coeff. Std.dev.     t
Constant  0.0000046  0.1275966     0.0000
X1       -0.2109841  0.1408017    -1.4984  Ikä
X2       -0.0718719  0.2547557    -0.2821  Pituus
X3       -0.4495819  0.1979440    -2.2713  Pituuskasvu
X4       -0.2024115  0.2739847    -0.7388  Tilavuuskasvu
Variance of regressand Y2=1.000057924 df=29
Residual variance=0.488427025 df=25
R=0.7609 R^2=0.5790 Durbin-Watson=2.150

MASK=XXXX--Y--P
REGDIAG KUUSET CUR+2

Regression diagnostics on data KUUSET: N=30
Regressand Y3        # of regressors=5 (Constant term included)
Condition number of scaled X: k=4.37129
Variable Regr.coeff. Std.dev.     t
Constant -0.0000195  0.1625030    -0.0001
X1        0.2290114  0.1793206     1.2771  Ikä
X2       -0.3573115  0.3244487    -1.1013  Pituus
X3        0.4939299  0.2520952     1.9593  Pituuskasvu
X4        0.2681527  0.3489382     0.7685  Tilavuuskasvu
Variance of regressand Y3=1.000064612 df=29
Residual variance=0.792216930 df=25
R=0.5631 R^2=0.3171 Durbin-Watson=2.158
```



Kaikkien regressiomallien yhteiskorrelaatiokertoimet (R) jäävät pienemmiksi kuin suurin kanoninen korrelaatio. Pituuskasvu on ainoa merkitsevä selittäjä. Kokonaiskuvan saaminen on hankalaa, sillä kuntoa kuvaavat muuttujat korreloivat keskenään:

```
MASK=----AAA---
CORR KUUSET

/LOADCORR
Limits: P=0.001 0.57   P=0.01 0.458   P=0.05 0.36
R(KUUSET)
             Y1      Y2      Y3
Y1        1.000   0.666  -0.343
Y2        0.666   1.000  -0.525
Y3       -0.343  -0.525   1.000
```

Tämäntyyppisissä tilanteissa kanoninen analyysi voikin olla ilmiön kuvaamisessa ja selittämisessä hyödyllisempi kuin regressioanalyysi.

SAS:issa kanoninen analyysi tapahtuu proseduurilla CANCORR.

Moniulotteinen skaalaus

Moniulotteinen skaalaus *(multidimensional scaling, MDS)* tarkoittaa havaintojen välisten etäisyyksien visualisoimista vähempiulotteisessa avaruudessa. Käytännössä tämä tarkoittaa useimmiten pyrkimistä tavalliseen kaksiulotteiseen diagrammiin, jonka toivotaan kuvaavan tilanteen riittävän tarkasti.

Analyysin lähtökohtana on etäisyys- tai erilaisuusmatriisi aivan vastaavasti kuin osassa ryhmittelymenetelmistä. Moniulotteinen skaalaus muistuttaa pääkomponenttianalyysia, eikä ihme, sillä ns. klassinen skaalaus palautuu etäisyysmatriisin tietyllä muunnoksella pääkomponenttianalyysiin.

Klassinen skaalaus ei käytännössä riitä, varsinkaan mikäli etäisyysmatriisi ei ole euklidinen tai sitä ei sellaiseksi saada muunnettua. Sillä saatua ratkaisua voi kuitenkin käyttää pohjana pienimmän neliösumman skaalaukselle *(least squares scaling)*, joka on periaatteeltaan yksinkertainen (minimoidaan havaittujen ja koordinaattien perusteella laskettujen etäisyyksien neliösummaa), tosin laskennallisesti raskas.

Aikoinaan (1960-luvulla) kehitettiin myös järjestysasteikolliselle mittaukselle sopivia ordinaaliskaalausmenettelyjä, mutta pienimmän neliösumman skaalaus todettiin paremmaksi jo 1980-luvulla. Vähän myöhemmin sen käytännön soveltaminenkin tuli mahdolliseksi, kun tietokoneisiin saatiin tarpeeksi tehoa.

Tarkastellaan aluksi erästä ekologisissa sovelluksissa tyypillistä mittaa, joka kulkee *Jaccardin indeksin* nimellä. Se on esimerkki binäärisistä etäisyysmitoista, joissa kiinnostaa vain onko jokin asia havaittu molemmissa vertailtavissa kohteissa, jommassa kummassa tai ei kummassakaan. Yleisesti tilanne voidaan kuvata seuraavanlaisena taulukkona (1=havaittu, 0=ei havaittu):

	1	0	
1	a	b	a+b
0	c	d	c+d
	a+c	b+d	N

(N=a+b+c+d)

Tästä voidaan muodostaa lukuisia erilaisia samankaltaisuus- tai erilaisuusmittoja. Jaccardin indeksi saadaan laskemalla sellaisten havaintoparien lukumäärä, joissa tutkittu asia on havaittu molemmissa ja jakamalla tämä sellaisten lukumäärällä, joissa asia on havaittu ainakin toisessa. Kaavamuodossa Jaccardin indeksi on siis a / (a+b+c), eli se jättää huomiotta havaintoparit joissa ei kummassakaan ole asiaa havaittu. Jaccardin indeksi on samankaltaisuusmitta, joten sen arvot on vielä muutettava kuvaamaan erilaisuuksia, esim. vähentämällä ne ykkösestä.

Sovelletaan nyt Jaccardin indeksiä pieneen metsäaineistoon. Esimerkki on Biometria-kirjasta. Aineistoon on taulukoitu Ahvenanmaan mantereen viideltä kuusivaltaiselta lehtometsäalueelta kerättyjen 25 maakiitäjäislajin runsaudet.

```
FILE STATUS METSÄT
 Lehtometsäaineisto / Biometria 3.painos s.504
FIELDS: (active)
   1 NA_   1 Laji     Maakiitäjäislajin numero
   2 NA_   2 X1       Lehtometsäalue 1
   3 NA_   1 X2       Lehtometsäalue 2
   4 NA_   1 X3       Lehtometsäalue 3
   5 NA_   2 X4       Lehtometsäalue 4
   6 NA_   1 X5       Lehtometsäalue 5
END
Survo data file METSÄT: record=30 bytes, M1=11 L=64  M=6 N=25
```

Tilan säästämiseksi aineisto on esitetty alla transponoituna.

```
Laji 1  2 3  4 5 6 7 8 9 10 11 12 13 14 15 16  17 18 19 20 21 22 23 24 25
X1   5  0 1  0 0 7 4 0 0 0  0  3  40 0  12 252 9  5  9  28 1  5  0  0  1
X2   3  0 0  1 0 2 8 5 0 0  0  6  19 3  4  97  12 0  1  5  2  7  1  2  0
X3   0  0 0  0 0 0 4 0 1 0  0  1  5  0  2  88  9  1  0  2  0  18 0  0  0
X4   3  6 0  0 1 0 3 2 2 0  2  1  49 0  2  106 3  1  3  8  2  4  0  0  0
X5   23 6 12 0 0 1 2 0 1 1  0  4  29 0  13 54  4  2  15 7  2  3  0  1  0
```

Lasketaan metsäalueiden väliset samankaltaisuudet Jaccardin indeksin avulla. Esimerkiksi metsien ja X1 ja X4 osalta a=12, b=3, c=5 ja d=5 (vaikkei sitä tarvitakaan), joten a / (a+b+c)=0.6 , mutta käsin laskeminen (josta 1980-luvulla kirjoitettu Biometria-kirja vielä muistuttaa) ei tietenkään ole enää relevanttia. Ohjelmat laskevat, tutkija voi keskittyä mielenkiintoisempiin tehtäviin. Muodostetaan siis koko etäisyysmatriisi (vähennetään jo samalla arvot ykkösestä).

```
MASK=-AAAAA
DISTV METSÄT,METSÄT / MEASURE=BINARY COEFF=1-a/(a+b+c)

MAT LOAD METSÄT
BINARY_1-a/(a+b+c)
///          X1     X2     X3     X4     X5
X1        0.000  0.400  0.438  0.400  0.263
X2        0.400  0.000  0.579  0.455  0.409
X3        0.438  0.579  0.000  0.412  0.444
X4        0.400  0.455  0.412  0.000  0.333
X5        0.263  0.409  0.444  0.333  0.000
```

Kuvataan nyt metsäalueiden väliset erilaisuudet moniulotteisen skaalauksen avulla. Aloitetaan tekemällä etäisyysmatriisille klassinen skaalaus kahteen ulottuvuuteen.

```
/CSCAL METSÄT,2
Classical multidimensional scaling for METSÄT:
Eigenvalues
///          DIM1     DIM2     DIM3     DIM4     DIM5
Eigenval 0.171968 0.079824 0.073817 0.027991 0.000000

Eigenvalues_(in_percentages)
///          DIM1     DIM2     DIM3     DIM4     DIM5
Per_cent  48.6335  22.5746  20.8760   7.9159   0.0000
Cumulat.  48.6335  71.2081  92.0841 100.0000 100.0000
```

Analyysi antaa tiedot etäisyysmatriisin ulottuvuuksista sen ominaisarvojen perusteella aivan vastaavasti kuin pääkomponenttianalyysissa. Lisäksi saadaan havaintopisteiden koordinaatit kaksiulotteisessa ratkaisuavaruudessa sekä näiden perusteella uudelleenlasketut etäisyydet.

```
CS_scales
///         DIM1   DIM2
X1         0.048 -0.180
X2         0.277  0.100
X3        -0.296  0.016
X4        -0.065  0.165
X5         0.036 -0.100

CS_distances
///           X1     X2     X3     X4     X5
X1         0.000  0.361  0.396  0.363  0.081
X2         0.361  0.000  0.579  0.348  0.313
X3         0.396  0.579  0.000  0.275  0.351
X4         0.363  0.348  0.275  0.000  0.284
X5         0.081  0.313  0.351  0.284  0.000
```

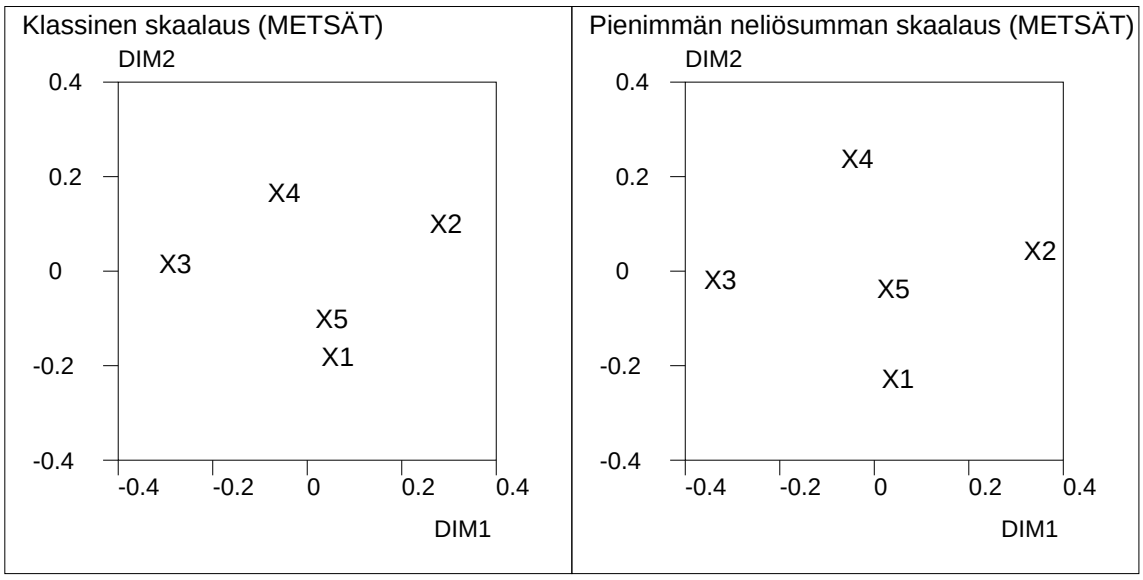
Etäisyyksissä näyttäisi olevan poikkeamia alkuperäisiin verrattuna. Tarkistetaan matriisitulkilla, paljonko poikkeamien neliöiden summa on:

```
MAT E=SUM(VEC(METSÄT-CSDIST.M),2)
MAT_E(1,1)=0.17657423506421
```

Tämä on juuri se, mitä pienimmän neliösumman skaalauksessa minimoidaan. Kokeillaan, paranisiko tulos sen avulla. Kysymyksessä on siis iteratiivinen ratkaisu, jonka pohjana hyödynnetään klassisen skaalauksen tulosta.

```
LSCAL METSÄT,CSCAL.M,END+2

Least-squares scaling for 5*5 dissimilarity (distance) matrix METSÄT:
Initial criterion value 0.176574  Dimension=2
Final criterion value 0.076092  nf=1104
```

Ratkaisu paranee jonkin verran, eli nyt saadu pistekonfiguraatio on lähempänä alkuperäisiä etäisyystietoja. On aika piirtää kuvia.



Ratkaisuissa ei ole oleellisia eroja, joten klassinen skaalauskin olisi riittänyt.

Korrelaatioihin pohjaavissa menetelmissä kuten faktorianalyysissa aineisto on miltei aina perinteinen havaintomatriisi: rivit kuvaavat havaintoja ja sarakkeet muuttujia. Ekologisissa aineistoissa ei ole useinkaan niin vaan rivit ja sarak-

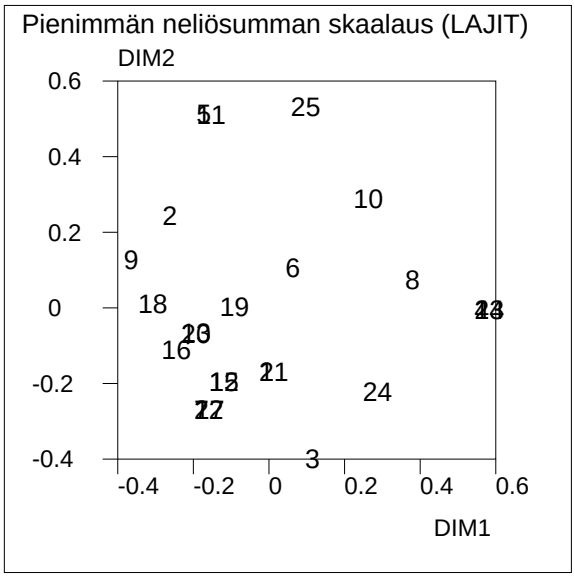
keet voivat hyvin olla symmetrisessä asemassa toisiinsa nähden. Niinpä etäisyyksiä voidaan laskea yhtä hyvin sarakkeiden kuin rivien välillä. Edellä esitellyssä aineistossa riveinä olivat havaintotiedot maakiitäjäisistä ja sarakkeina metsät. Metsien väliset etäisyydet maakiitäjäisten suhteen laskettiin siis aineiston sarakkeiden välisinä etäisyyksinä. Toisinpäin käännettynä tämä tarkoittaisi lajien välisiä eroja metsien suhteen. Tässä se saattaa olla hieman keinotekoista, mutta katsotaan silti miltä tarkastelu näyttäisi niin päin. Ei tässä kovin pahasti metsään mennä.

```
DIST METSÄT,LAJIT / MEASURE=BINARY COEFF=1-a/(a+b+c)
```

Lajien etäisyysmatriisi on kooltaan 25x25. Klassinen skaalaus ilmoittaa ettei matriisi ole euklidinen, joten kannattaa jatkaa samantien pienimmän neliösumman skaalaukseen.

```
LSCAL LAJIT,CSCAL.M,END+2

Least-squares scaling for 25*25 dissimilarity (distance) matrix LAJIT:
Initial criterion value 18.6788  Dimension=2
Final criterion value 8.7168  nf=6636
```

Minimoitava neliösumma on pienentynyt aika selvästi (yli 6000 iteroinnin jälkeen). On muistettava että tässä matriisissa on 25*25=625 alkiota, kun edellisessä oli vain 5*5=25.



Lajit ryhmittyvät vähän kuin ympyrän muotoon, mikä on aivan tyypillistä. En osaa sanoa, olisiko tästä löydettävissä jotain tulkintaa. Olisi ainakin tunnettava substanssialan teoriaa enemmän. Joka tapauksessa teknisesti on samantekevää miten päin tällaisia aineistoja tutkitaan. Järkevät kysymyksenasettelut ovat tietenkin tutkijan vastuulla, kuten yleensäkin.

Moniulotteinen skaalaus on kätevä kuvaustapa, mutta se yltää vain yhden asian skaalaamiseen kerrallaan. Jos tarkasteluja halutaan tehdä yhtaikaa rivien ja sarakkeiden suunnassa, on siirryttävä korrespondenssianalyysiin.

SAS:in skaalausproseduuri on nimeltään MDS.

Korrespondenssianalyysi

Korrespondenssianalyysi *(correspondence analysis)* voidaan lukea yhtä hyvin monimuuttujamenetelmien kuin frekvenssiaineistojen analyysimenetelmien joukkoon. Sillä on erittäin värikäs historia, ja se on tunnettu aikojen saatossa lukuisilla eri nimillä, mm. *optimal scaling*, *reciprocal averaging*, *optimal scoring*, *appropriate scoring*, *quantification method*, *homogeneity analysis*, *dual scaling* ja *scalogram analysis*.

Menetelmän perusidea on yksinkertainen: visualisoidaan taulukkomuotoisen aineiston sisältämä informaatio niin että mm. nähdään miten sarakkeiden ja rivien tiedot suhtautuvat toisiinsa, eli mitä vastaavuuksia *(correspondence)* niihin sisältyy. Tästä teemasta on lukuisia muunnelmia, mutta päämäärä on kaikissa sama: aineiston kuvaaminen. Se on toki muillekin monimuuttujamenetelmille ominaista, mutta tässä menetelmässä kuvanpiirron merkitys korostuu erityisesti. Korrespondenssianalyysin yhteydessä on luonnollista soveltaa 1970-luvun alussa kehitettyä ns. *biplot*-tekniikkaa, jossa samaan kuvaan yhdistetään lukuisia erilaisia asioita, jopa eri koordinaatistoja päällekkäin. Kuvat vaativat helposti jonkin verran lukutaitoa auetakseen, mutta tarjoavat osaavalle monipuolisia näkymiä aineistoon ja edelleen tutkittavaan ilmiöön.

Nykymuotoonsa korrespondenssianalyysin kehittivät Ranskassa mm. *Jean-Paul Benzécri* ja *Ludovic Lebart* 1970-luvulla nimellä *analyse factorielle des correspondances*, mutta jo 1930- ja 40-luvuilla menetelmää kehittelivät toisistaan riippumatta mm. *H. Hirschfeld* ja *R. A. Fisher*, sekä myöhemmin *Chikio Hayashi*, joka teki menetelmän tunnetuksi Japanissa. 1970-luvulla korrespondenssianalyysi oli anglosaksisissa maissa likipitäen unohdettu, lukuunottamatta joitakin *M. Hillin* yrityksiä herättää asia jälleen henkiin. Kuitenkin vasta *Michael Greenacren* 1980-luvulla julkaisema kirja ja sitä seuranneet muut oppikirjat korrespondenssianalyysista ovat nostaneet menetelmän tunnettuutta ja suosiota vuosi vuodelta yhä useammilla aloilla.

Eräs analyysin muunnelmista tunnetaan nimellä moniulotteinen korrespondenssianalyysi *(multiple correspondence analysis, MCA)*. Asiallisempi suomennos olisi *moninkertainen*, sillä kyse on edelleen kaksiulotteisista taulukoista. Tässä ne on vain koottu yhteen matriisiksi, jota kutsutaan *Burtin* tauluksi. Näin päästään visualisoimaan useamman kuin kahden asian välisiä yhteyksiä. *Cyril Burt* luetaan siis myös tämän menetelmän laajaan kehittäjäjoukkoon, sillä hän keksi ko. tauluesityksen 1950-luvun alussa pohtiessaan mahdollisuutta kvalitatiivisten muuttujien *faktorianalyysiin* (vrt. ranskankielinen alkuperäinen nimitys edellä). Kuulemma Ranskassa saatetaan nykyisinkin puhua faktorianalyysista tarkoittaen korrespondenssianalyysia.

Ekologian tutkimuksessa korrespondenssianalyysi on kehittynyt moniulotteisen skaalauksen suunnalta *(dual scaling)*. Lähtökohtana on paikkojen *(sites)* ja lajien *(species)* välinen lajilukumäärien tms. taulukko, ja mielenkiinto kohdistuu siihen, millaisia ordinaatioita, ryhmityksiä, trendejä, gradientteja ym. vastaavuuksia näistä muodostuu. Tärkeässä osassa ovat näiden ohella ympäristömuuttujat, jotka ovat paikkojen tilaa kuvaavia mittauksia. Myös lajeihin voi liittyä taustatekijöitä. Ympäristö- ym. taustatekijöiden vaikutus halutaan saada esille samassa analyysissa.

Pohjana olivat 1960-luvulla kehitetyt ordinaaliskaalausmenettelyt sekä *R. H. Whittakerin* gradienttianalyysina mm. kasviekologiassa tunnettu yksinkertaistettu skaalaus. Kytkennän korrespondenssianalyysiin loi 1970-luvulla *M. Hill*, joka kehitti menetelmästä ns. kaariefektin *(arch effect)* eliminoivan, mutta myöhemmin kiistellyn muunnelman nimeltä oikaistu korrespondenssianalyysi *(detrended correspondence analysis, DCA)* sekä DECORANA-nimisen ohjelman sen soveltamiseen.

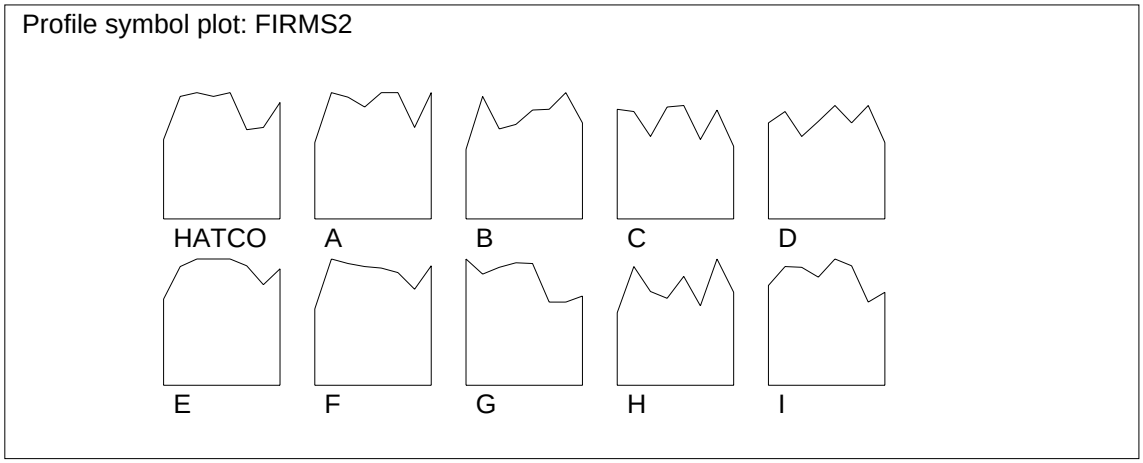
Varsinaisesti menetelmän teki ekologiassa tunnetuksi 1980-luvulla hollantilainen *Cajo J. F. ter Braak*, joka kehitti ns. kanonisen korrespondenssianalyysin *(canonical correspondence analysis, CCA)* sekä laati sitä varten CANOCO-nimisen ohjelman. Tässä muunnelmassa taustatekijöiden riippuvuudet laji- ja paikkatiedoista oletetaan lineaarisiksi, jolloin ne voidaan ratkaista regressioanalyysin avulla.

Tavallinen korrespondenssianalyysi

Tarkastellaan aluksi pientä esimerkkiä markkinatutkimuksen alalta, jossa tavallisella korrespondenssianalyysilla on ollut yhä runsaammin käyttöä viime vuosina. Aineisto on Hairin ym. kirjassa esitetystä kuvitteellisen HATCO-yhtiön *(Hair Anderson Tatham COmpany)* tietokannasta tehty frekvenssitaulukko, jossa riveinä ovat yritysten ominaisuuksia kuvaavat (binääriset) attribuutit ja sarakkeina yritykset (HATCO sekä sen kilpailijat A, B, ..., I). Luvut kertovat joka attribuutin kohdalla, kuinka moni kyselyyn vastanneista on liittänyt ko. attribuutin kuhunkin yritykseen.

Attribute	HATCO	A	B	C	D	E	F	G	H	I	Sum
Product quality	4	3	1	13	9	6	3	18	2	10	69
Strategic orientation	15	16	15	11	11	14	16	12	14	14	138
Overall service	15	14	6	4	4	15	14	13	7	13	105
Delivery speed	16	13	8	13	9	17	15	16	6	12	125
Price level	14	14	10	11	11	14	12	13	10	14	123
Salesforce image	7	18	13	4	9	16	14	5	4	16	106
Price flexibility	6	6	14	10	11	8	7	4	14	4	84
Manufacturer image	15	18	9	2	3	15	16	7	8	8	101
Sum	92	102	76	68	67	105	97	88	65	91	851

Kun silmäilee lukuja, havaitsee että yrityksillä on erilaisia *profiileja* attribuuttien suhteen. Helpommin sen havaitsee oheisesta ns. profiilikuvasta. HATCOn profiili muistuttaa eniten yrityksen E profiilia. Myös F ja I ovat samantyyppisiä, kun taas esimerkiksi C näyttää aivan erilaiselta.

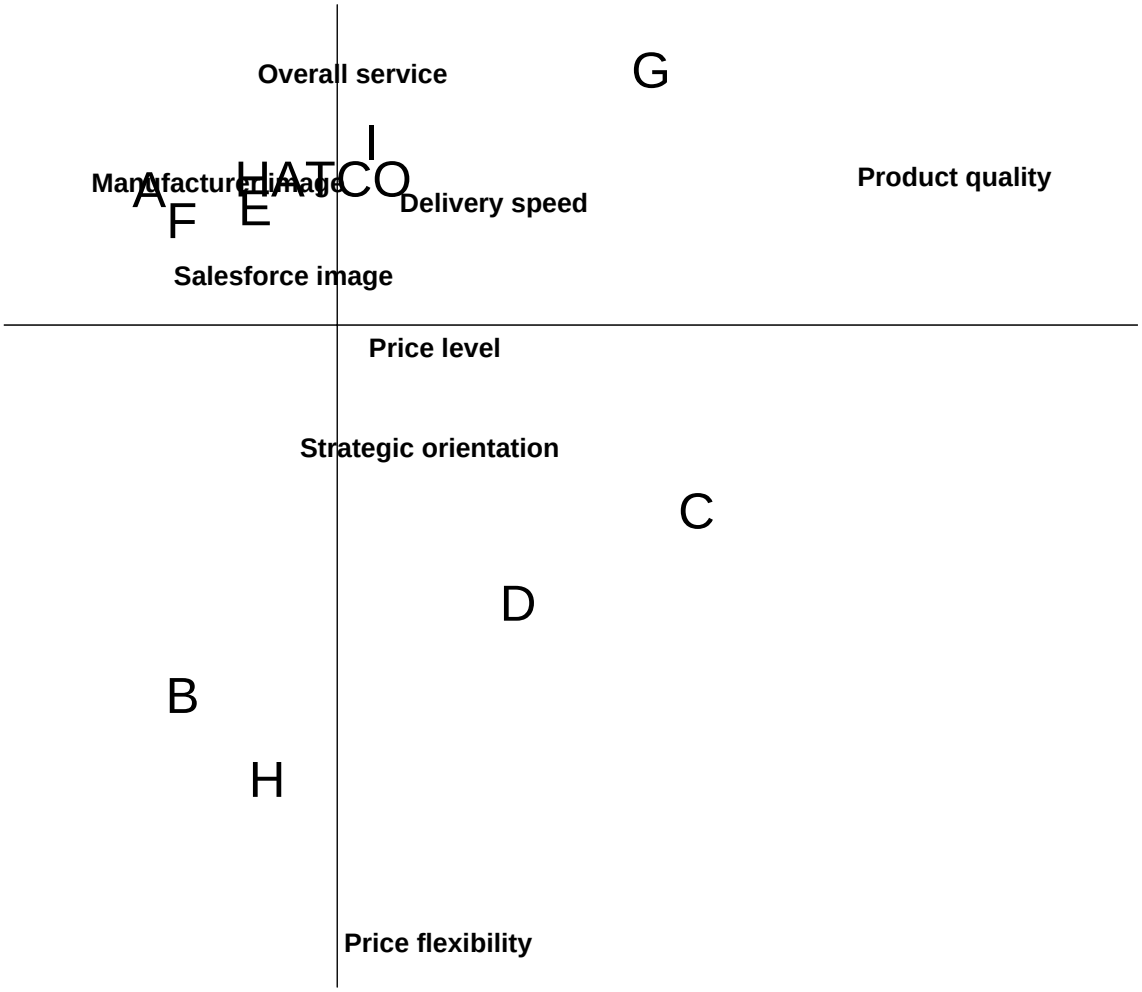


Mutta mitä profiilit piilottavat taakseen? Myös attribuuttien tarkempi tyypittely tai ryhmittely olisi kiinnostavaa. Siirrytään siis korrespondenssianalyysiin, joka huolehtii taulukon molemmista suunnista samanaikaisesti.

```
CORRESP FIRMS,CUR+2

Correspondence analysis on data FIRMS: Rows=8 Columns=10

    Canonical      Eigen-           Chi^2    Cumulative
    correlation    value                     percentage
 1  0.2767         0.0765      65.1367748       53.13
 2  0.2187         0.0478      40.6891946       86.32
 3  0.1237         0.0153      13.0130652       96.93
 4  0.0516         0.0027      2.26158532       98.78
 5  0.0284         0.0008      0.68562546       99.34
 ...
                   0.1441     122.601 (df=63 P=1.03329e-005)
```

Analyysin yleistuloksista voidaan päätellä, että riittänee käsitellä kahta ensimmäistä dimensiota, jotka yhteensä selittävät ilmiön vaihtelusta n. 53+33=86 %. Muiden monimuuttujamenetelmien tapaan tässäkin esiintyy kanonisia korrelaatioita. Niillä tarkoitetaan taulukon luokittelijoiden (yritykset ja attribuutit) skaalausten välisiä maksimaalisia korrelaatioita. Suurin on vain luokkaa 0.28, mutta se ei kerro läheskään kaikkea, eikä ole tarkoituskaan.

Piirretään analyysin päätulos eli rivien ja sarakkeiden skaalaus samaan kuvaan. Koordinaattiakseleina ovat siis dimensiot 1 ja 2. Yksiköt jätetään useimmiten pois, koska vain suhteelliset erot kiinnostavat.



Tulostaulukoista nähdään, että ensimmäisen dimension vaihtelusta n. 86 % koostuu attribuuteista *Product quality* ja *Manufacturer image*. Vastaavasti toisella dimensiolla painottuvat *Price flexibility* ja *Overall service* (yht. 83 %).

```
Absolute contributions %

Attribute                  Abs1   Abs2
Product quality           66.48   3.29
Strategic orientation      0.38   4.47
Overall service            1.82  14.23
Delivery speed             2.22   4.04
Price level                0.69   0.14
Salesforce image           8.72   0.55
Price flexibility          0.07  68.86
Manufacturer image        19.61   4.42
Sum                      100.00 100.00
```

Yritysten osalta tilanne on jonkin verran tasaisempi, mutta silti C ja G yhdessä vastaavat n. 60 % vaihtelusta ensimmäisellä, B ja H puolestaan n. 50 % toisella dimensiolla. HATCOn osuus on hyvin pieni molemmilla dimensioilla.

```
Firm                       Abs1   Abs2
HATCO                      2.39   4.24
A                         12.51   4.03
B                          6.35  22.39
C                         29.87   5.03
D                          7.41  11.13
E                          2.50   3.11
F                          8.01   2.28
G                         29.21  12.33
H                          1.18  28.88
I                          0.58   6.58
Sum                      100.00 100.00
```

Toisinpäin tarkasteluna, eli miten suuren osan kunkin muuttujan varianssista dimensiot selittävät, parhaat attribuutit ovat *Product quality* ja *Price flexibility*. Vain *Salesforce image* jää kahdella dimensiolla yhteensä alle 0.5:n.

```
Squared correlations

Attribute                  Sqr1   Sqr2   Sum
Product quality           0.961  0.030  0.991
Strategic orientation     0.093  0.678  0.771
Overall service           0.138  0.677  0.815
Delivery speed            0.289  0.330  0.619
Price level               0.469  0.058  0.527
Salesforce image          0.358  0.014  0.372
Price flexibility         0.002  0.989  0.991
Manufacturer image        0.789  0.111  0.900
```

Yrityksistä puolet pärjää oikein hyvin. Vain HATCO ja I jäävät alle 0.5:n.

```
Firm                       Sqr1   Sqr2  Sum
HATCO                     0.206  0.228  0.434
A                         0.772  0.156  0.928
B                         0.294  0.648  0.942
C                         0.882  0.093  0.975
D                         0.445  0.418  0.863
E                         0.456  0.356  0.812
F                         0.810  0.144  0.954
G                         0.762  0.201  0.963
H                         0.049  0.748  0.797
I                         0.055  0.390  0.445
```

Tulkinnan apuna voi lisäksi käyttää rivien ja sarakkeiden painoja eli niiden suhteellisia osuuksia. Tässä kaikki ovat aika tasavahvoja, attribuuteista heikoin on *Product quality* ja vahvin *Strategic orientation*.

```
Masses (relative frequencies):

Attribute                 Mass
Product quality           0.081   69/851=0.08108108108108
Strategic orientation     0.162  138/851=0.16216216216216
Overall service           0.123
Delivery speed            0.147
Price level               0.145
Salesforce image          0.125
Price flexibility         0.099
Manufacturer image        0.119
Sum                       1.000
```

Yrityksistä heikoin on H ja vahvin E. Nämä luvut saadaan siis suoraan edellä olevan taulukon rivi- ja sarakesummien avulla.

```
Firm                       Mass
HATCO                     0.108
A                         0.120
B                         0.089
C                         0.080
D                         0.079
E                         0.123  105/851=0.12338425381904
F                         0.114
G                         0.103
H                         0.076   65/851=0.07638072855464
I                         0.107
Sum                       1.000
```

Kuvan ja tulostaulukoiden avulla selviää mm., että tuotteen laatu *(Product quality)* liitetään eniten yrityksiin G ja C, ja joustavat hinnat *(Price flexibility)* yrityksiin H, B, D ja C. Yritys C pärjää siis hyvin molemmilla dimensioilla.

Viimeksi mainittuja neljää pidetään huonoimpina palvelun *(Overall service)* suhteen. B on kaikista huonoin laadun suhteen. HATCOa arvostetaan monessa suhteessa, esimerkiksi toimitusnopeudessa *(Delivery speed)*, mutta siinä ovat useimmat muutkin hyviä joten sekä HATCOn että ao. attribuutin pisteet jäävät kuvassa lähemmäksi origoa. Parhaiten selittävien attribuuttien osalta HATCO ei ole menestynyt. Tulkintaa voidaan tästä edelleen jatkaa ja tarkentaa. Kaikki nämä asiat voidaan nähdä myös suoraan alkuperäisen taulukon 80 lukua katselemalla, mutta yksi kuva välittää varmasti saman informaation nopeammin.

Tästä ei korrespondenssianalyysi oleellisesti muuksi muutu, vaikka erilaisia muunnelmia onkin olemassa melkoinen liuta. Katsotaan paria niistä lyhyesti.

Moniulotteinen korrespondenssianalyysi

Tarkastellaan moniulotteista (moninkertaista) korrespondenssianalyysia pienen maatilaesimerkin avulla.

Farm number	Moisture class	Grassland management	Grassland use	Manure class
1	1	SF	2	4
2	1	BF	2	2
3	2	SF	2	4
4	2	SF	2	4
5	1	HF	1	2
6	1	HF	2	2
7	1	HF	3	3
8	5	HF	3	3
9	4	HF	1	1
10	2	BF	1	1
11	1	BF	3	1
12	4	SF	2	2
13	5	SF	2	3
14	5	NM	3	0
15	5	NM	2	0
16	5	SF	3	3
17	2	NM	1	0
18	1	NM	1	0
19	5	NM	1	0
20	5	NM	1	0

Data observed at 20 farms on the island of Terschelling (from Jongman, ter Braak and van Tongeren, 1987), example taken from Gower, J. C., & Hand, D. J. (1996). Biplots. Chapman & Hall, London.

Muuttujat ovat siis kosteus *(moisture)*, jossa on neljä luokkaa (M1, M2, M4, M5), viljelytyyppi, myös neljä luokkaa: *standard farming* (SF), *biological farming* (BF), *hobby farming* (HF) ja *nature conservation management* (NM). Muut kaksi muuttujaa ovat maankäyttö: *hay production* eli heinäntuotanto (U1), *intermediate* eli sekä että (U2) ja *grazing* eli laidunmaa (U3) sekä lannoitus *(manure)*, jolla on järjestysasteikollinen luokitus C0, ..., C4.

Muodostetaan ns. Burtin taulu. Se on symmetrinen matriisi, jonka keskellä (tässä korostettuina) ovat lävistäjämatriisit havaintojen lukumääristä luokittain ja muualla kaikki muuttujien parittaiset ristiintaulukot.

Burt's_table_of_DATA_FARMS

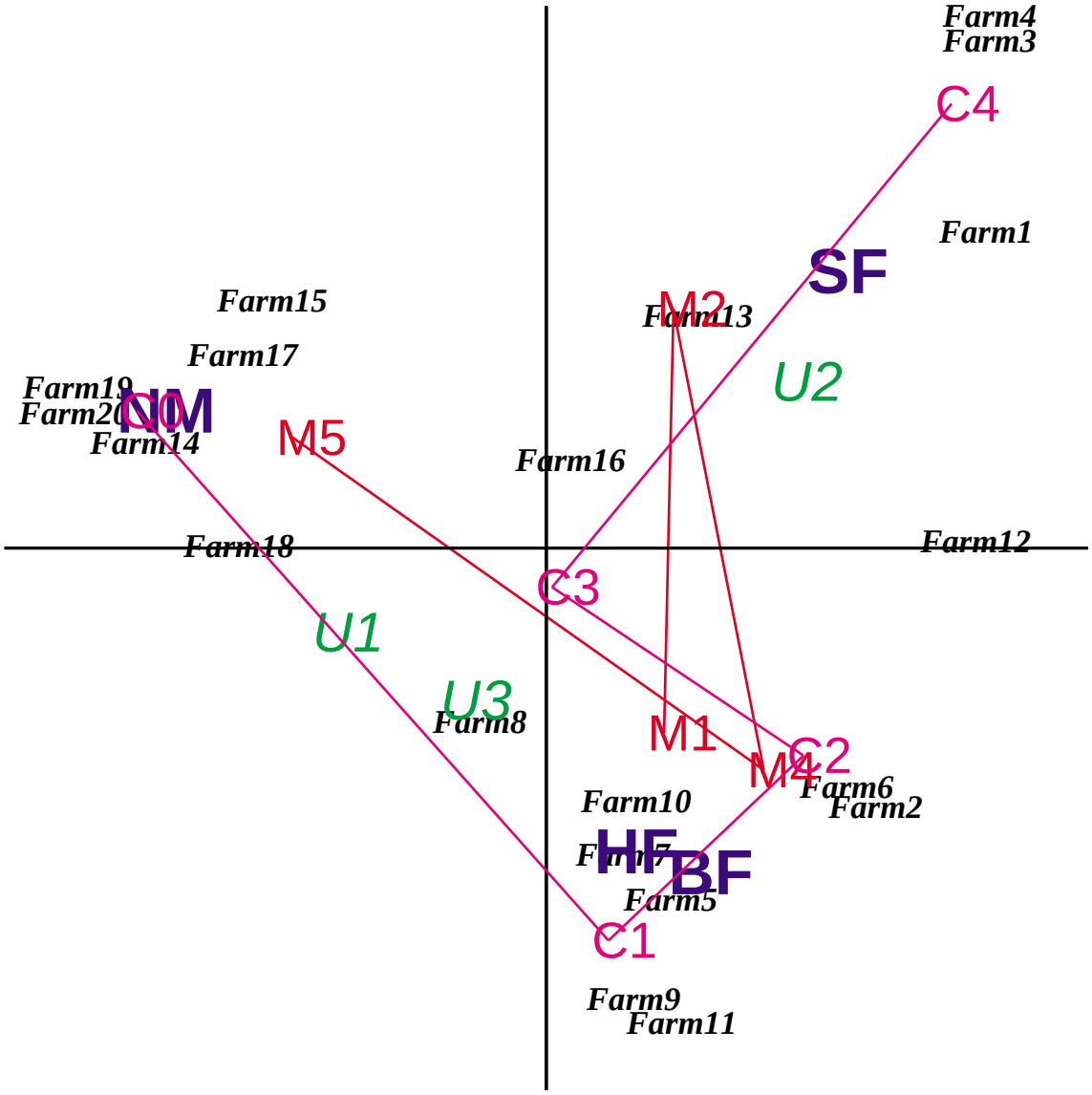
	M1	M2	M4	M5	SF	BF	HF	NM	U1	U2	U3	C0	C1	C2	C3	C4
M1	**7**	**0**	**0**	**0**	1	2	3	1	2	3	2	1	1	3	1	1
M2	**0**	**4**	**0**	**0**	2	1	0	1	2	2	0	1	1	0	0	2
M4	**0**	**0**	**2**	**0**	1	0	1	0	1	1	0	0	1	1	0	0
M5	**0**	**0**	**0**	**7**	2	0	1	4	2	2	3	4	0	0	3	0
SF	1	2	1	2	**6**	**0**	**0**	**0**	0	5	1	0	0	1	2	3
BF	2	1	0	0	**0**	**3**	**0**	**0**	1	1	1	0	2	1	0	0
HF	3	0	1	1	**0**	**0**	**5**	**0**	2	1	2	0	1	2	2	0
NM	1	1	0	4	**0**	**0**	**0**	**6**	4	1	1	6	0	0	0	0
U1	2	2	1	2	0	1	2	4	**7**	**0**	**0**	4	2	1	0	0
U2	3	2	1	2	5	1	1	1	**0**	**8**	**0**	1	0	3	1	3
U3	2	0	0	3	1	1	2	1	**0**	**0**	**5**	1	1	0	3	0
C0	1	1	0	4	0	0	0	6	4	1	1	**6**	**0**	**0**	**0**	**0**
C1	1	1	1	0	0	2	1	0	2	0	1	**0**	**3**	**0**	**0**	**0**
C2	3	0	1	0	1	1	2	0	1	3	0	**0**	**0**	**4**	**0**	**0**
C3	1	0	0	3	2	0	2	0	0	1	3	**0**	**0**	**0**	**4**	**0**
C4	1	2	0	0	3	0	0	0	0	3	0	**0**	**0**	**0**	**0**	**3**

Moniulotteinen korrespondenssianalyysi tehdään yleensä Burtin taulusta, koska se on hyvin kompakti datan esitysmuoto. Toinen mahdollisuus on tehdä analyysi binäärisestä indikaattorimatriisista, joka tässä tapauksessa ei ole juuri Burtin taulua suurempi (koska maatiloja on vain 20). Yleensä se on huomattavasti suurempi, eikä sen käsittely ole lainkaan kätevää.

Binary_form_of_DATA_FARMS

	M1	M2	M4	M5	SF	BF	HF	NM	U1	U2	U3	C0	C1	C2	C3	C4
1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1
2	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0
3	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	1
4	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	1
5	1	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0
6	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0
7	1	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0
8	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0
9	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0
10	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0
11	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0
12	0	0	1	0	1	0	0	0	0	1	0	0	0	1	0	0
13	0	0	0	1	1	0	0	0	0	1	0	0	0	0	1	0
14	0	0	0	1	0	0	0	1	0	0	1	1	0	0	0	0
15	0	0	0	1	0	0	0	1	0	1	0	1	0	0	0	0
16	0	0	0	1	1	0	0	0	0	0	1	0	0	0	1	0
17	0	1	0	0	0	0	0	1	1	0	0	1	0	0	0	0
18	1	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0
19	0	0	0	1	0	0	0	1	1	0	0	1	0	0	0	0
20	0	0	0	1	0	0	0	1	1	0	0	1	0	0	0	0

Aineistoon on sisällytetty myös maatilojen koodit omana luokittelevana muuttujanaan. Näin saadaan kuva, jossa tälle analyysille tyypillisesti on yhdistetty vähintään järjestettyjen muuttujien kategoriat toisiinsa viivoilla. Kuvasta voidaan suoraan lukea, mitkä maatilat ovat minkäkin luokituksen mukaisia, miten eri luokitukset suhtautuvat toisiinsa jne.



Esimerkiksi biologinen ja harrastusviljely (HF ja BF) menevät niin lähekkäin kuvassa, että niiden profiilit muiden asioiden suhteen ovat hyvin samankaltaiset. Niissä lannoitetaan vähän (C1, C2), ja kosteusluokitus on joko M1 tai M4 (hyvin kuiva tai melko märkä). Luomutilat (NM) taas ovat kosteimmilla (M5) alueilla, ja niillä ei lannoiteta lainkaan (C0). Maankäytön suhteen esiintyy lähes kaikkia yhdistelmiä, niinpä kategoriat U1, U2 ja U3 asettuvat tilatyyppien välimaastoon. Kaikkiaan tilat ryhmittyvät melko selkeästi kolmeen ryhmään.

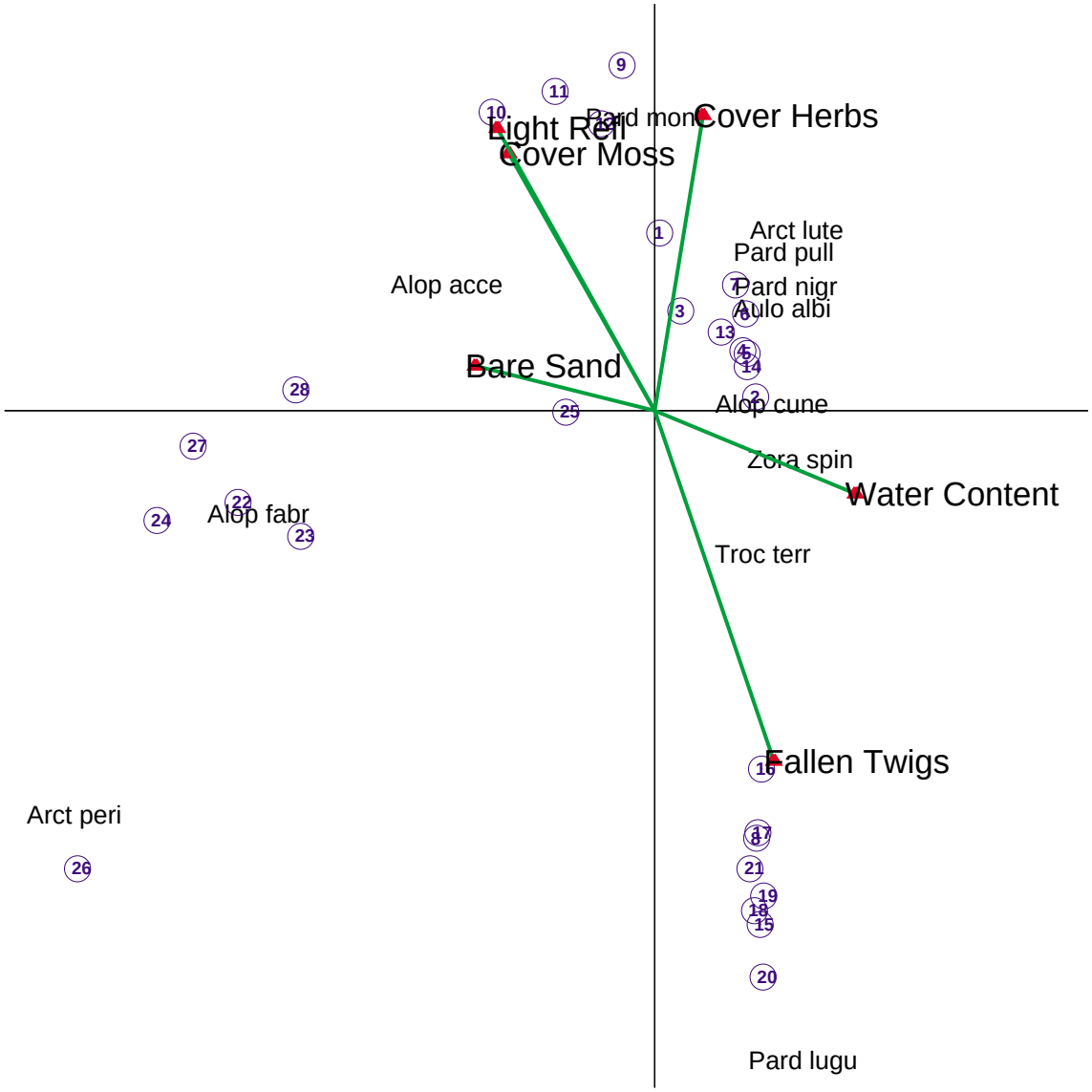
Kanoninen korrespondenssianalyysi

Tutkitaan lopuksi kanonista korrespondenssianalyysia sille tyypillisen esimerkin avulla. Aineisto koostuu lajeista ja paikoista, tässä eri hämähäkkilajeista Hollannin hiekkadyyneillä, sekä paikkoihin liittyvistä ympäristömuuttujista. Lajien runsaustiedoille *(abundance)* on tehty neliöjuurimuunnos, ja otettu siitä vain kokonaisosa. Korrespondenssianalyysi soveltuu siis lukumäärien ohella minkä tahansa ei-negatiivisten lukujen käsittelyyn. Ympäristömuuttujat on puolestaan ilmaistu karkeasti prosenttiosuuksina.

ter Braak: Hunting spiders in a Dutch dune area (Ecology, Vol. 67, No.5)

Species	**15**	19	**20**	16	**17**	18	**2**	8	**21**	5	**6**	14	**4**	7	**13**	3	**1**	9	**12**	25	**11**	10	**28**	23	**22**	27	**24**	26
Arct lute	0	0	0	0	0	0	0	0	0	1	2	1	1	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Pard lugu	2	3	3	2	1	2	1	7	4	1	0	1	1	1	1	1	0	0	0	1	0	0	0	1	0	0	0	0
Zora spin	1	1	1	2	1	0	3	1	1	4	5	5	5	4	4	1	2	0	0	2	0	0	0	0	0	0	0	0
Pard nigr	0	1	0	1	0	0	3	1	0	9	5	3	5	9	7	4	3	1	1	2	0	0	0	0	0	0	0	0
Pard pull	0	0	0	0	0	0	6	1	1	8	4	8	9	9	8	6	6	1	2	0	1	0	0	0	0	0	0	0
Aulo albi	0	0	0	0	0	0	5	2	0	3	2	2	4	4	4	3	2	0	0	1	1	0	0	0	0	0	0	0
Troc terr	5	4	4	5	4	5	8	5	4	9	7	9	9	9	9	8	7	1	3	4	2	1	1	1	1	0	0	1
Alop cune	0	1	1	1	0	1	1	3	1	4	2	1	2	2	6	4	3	1	3	1	1	0	0	0	0	0	0	0
Pard mont	0	0	0	0	0	0	1	1	1	1	3	3	2	5	4	5	7	5	9	3	9	4	2	2	1	1	1	0
Alop acce	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	3	5	1	4	3	3	1	3	4	2	5	3	1
Alop fabr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	3	1	1	3	3	4	3	4	2
Arct peri	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1	2	2	4
Water Content	9	7	8	8	9	8	8	6	7	8	9	8	6	8	9	6	5	5	5	3	4	4	0	0	1	0	2	0
Bare Sand	0	0	0	0	0	0	0	0	0	0	5	0	0	0	3	0	0	0	0	7	0	8	7	6	7	5	7	9
Cover Moss	1	3	1	1	1	0	2	2	1	0	5	4	5	1	1	5	7	9	8	2	9	7	8	9	9	8	9	4
Light Refl	1	0	0	0	2	2	3	1	0	5	1	2	6	5	7	8	8	7	8	5	8	8	8	9	8	8	9	9
Fallen Twigs	9	9	9	9	9	9	3	9	9	0	7	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
Cover Herbs	5	2	0	0	5	5	9	6	2	9	6	9	9	9	9	9	9	6	8	8	7	5	6	6	0	6	5	2

Paikkojen ja lajien väliset suhteet selviävät tavallisella korrespondenssianalyysilla. Ympäristömuuttujien osalta riittää laskea korrelaatiot paikkojen koordinaattipisteiden kanssa. Näin saatu oheinen kuva ei aivan tarkalleen vastaa alkuperäisen artikkelin esitystä, mutta samat johtopäätökset siitä on tehtävissä. Ympäristömuuttujat on tapana piirtää origosta lähtevinä vektoreina (vrt. erotteluanalyysi), jotka ilmentävät erilaisia jatkumoita suunnasta toiseen. Kuvan tulkinnassa noudatetaan aivan samoja periaatteita kuin edellä.



SAS:issa tavallisia ja moniulotteisia korrespondenssianalyyseja voi tehdä CORRESP-proseduurilla.