Monimuuttujamenetelmät

Erkki Liski
Matematiikan, Tilastotieteen ja Filosofian Laitos
Tampereen Yliopisto

1.1.2003

Sisältö

Luku 1

Johdanto

1.1 Miksi monimuuttujamenetelmät?

Monimuuttujainen aineisto syntyy, kun jokaisesta tilastoyksiköstä tarkastellaan useita eri ominaisuuksia samanaikaisesti. Aineistossa on siis tyypillisesti useita muuttujia, jotka on saatu yhdestä tai useammasta otoksesta tai kokeesta. Perusopinnoissa jokainen on taatusti jo tutustunut monimuuttujaiseen ainistoon, koska tilastolliset aineistot ovat pääsääntöisesti monimuuttujaisia. Esimerkiksi muuttujien välisten korrelaatioiden samanaikaista tarkastelua voi jo pitää monimuuttujaisena analyysina. Toisaalta tavallista regressioanalyysia ei pidetä varsinaisena monimuuttujamenetelmänä, vaikka selittäviä muuttujia olisikin useita. Monimuuttujaisessa regressiossa on myös useita selitettäviä muuttujia.

Jotkin monimuuttujaisen anyysin kysymykset ovat vastaavien yhden muuttujan ongelmien suoria yleistyksiä, kun taas toiset esiintyvät vain monimuuttujaisissa tutkimusasetelmissa. Samaa voidaan sanoa myös itse monimuuttujaiseen analyysiin käytettävistä menetelmistä. Esimerkiksi tavallinen t-testi voidaan voidaan suoraviivaisesti yleistää monimuuttujaiseksi testiksi. Sen sijaan pääkomponenttien määrittäminen voi tulla kyseeseen vain monimuuttujaisessa aineistossa.

1.2 Tarvittavat esitiedot

Kurssilla edellytetään, että matriisilaskennan perusteet ovat hallinnassa. Esimerkiksi kurssi *Matriisilaskentaa tilastotieteilijöille* antaa riittävät perusvalmiudet. Laskentaan kurssilla käytetään `R-ohjelmistoa`. Monimuuttulamenetelmien kannalta hyödyllisiä `R-kirjastoja` ovat esimerkiksi `MASS`, `Mva`, `Mvtnorm` ja `Multiv`.

1.3 Joitakin monimuuttujaisia ongelmia

1. Tuttujen yhden muuttujan analyysien yleistäminen monimuuttujaisiksi. Esimerkiksi t-testi, regressioanalyysi ja varianssianalyysi.

2. Aineiston moniulotteisuuden pienentäminen. Jos aineistossa on esimerkiksi joukko voimakkaasti keskenään korreloituneita muuttujia, voidaan osa niistä ehkä poistaa menettämättä oleellisesti informaatiota tai ne voidaan esittää muutaman muun muuttujan avulla. Esimerkiksi pääkomponenntianalyysi ja faktorianalyysi ovat tällaisia menetelmiä.

3. Erotteluanalyysi. Tietyn potilasjoukon epäillään sairastavan tautia A. Heille tehdään joukko laboratoriokokeita. Kokeiden tulosten perusteella pyritään muodostamaan diagnostinen sääntö, joka mahdollisimman hyvin erottelee tautia A sairastavat. Erotteluanalyysi tarjoaa keinoja hyvien erottelusääntöjen muodostamiseksi.

4. Ryhmittelyanalyysi/luokittelu

5. Kanooniset korrelaatiot.

Luku 2

Monimuuttujainen havaintoaineisto

2.1 Keskiarvovektorit

Olkoon otoksessa n yksilöä ja jokaisesta yksilöstä on mitattu muuttujat $y_1, y_2, \ldots, y_p$. *Havaintovektoreita* merkitään $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, missä i. havaintovektori on

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ip} \end{pmatrix}.$$

Aineiston *havaintomatriisin*

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_i' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1j} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nj} & \cdots & y_{np} \end{pmatrix}$$

riveinä ovat havaintovektorit transponoituina. Ensimmäinen alaindeksi i viittaa tilastoyksikköön ja toinen indeksi j viittaa muuttujaan: y_{ij} on i. tilastoyksikön mittaluku j. muuttujalla.

Muuttujan y_j *otoskeskiarvo* on $\bar{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$. Esimerkiksi $\bar{y}_1$ on n:n ensimmäisen muuttujan havaintoarvojen aritmeettinen keskiarvo, $\bar{y}_2$ on toisen muuttujan keskiarvo ja niin edelleen. Muuttujien $y_1, y_2, \ldots, y_p$ otoskeskiarvo on keskiarvovektori

$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_p \end{pmatrix}.$$

Taulukko 2.1 FIRMAT-aineisto. Yhdysvaltojen 10 suurimman teollisuusyrityksen tulokset vuodelta 1978

Yritys	y_1=varat*	y_2=nettotulot*	y_3=pörssiarvo*
G.M.	26.7	3.3	15.8
Exxon	38.4	2.4	19.5
Ford	19.2	1.7	8.4
Mobil	20.6	1.0	8.2
Texaco	18.9	0.9	9.4
Std. Oil	14.8	1.0	7.6
IBM	19.0	2.7	12.6
Gulf	14.2	0.8	7.3
G.E.	13.7	1.1	5.9
Chrysler	7.7	0.2	2.9

*miljoonia dollareita
Lähde: "Fortune 500", Fortune, 97, 238-272, May 8, 1978.

Matriisimerkinnöin otoskeskiarvovektori voidaan lausua

$$\bar{\mathbf{y}} = \frac{1}{n}\mathbf{Y}'\mathbf{1}_n, \tag{2.1.1}$$

missä $\mathbf{1}_n = (1, 1, \ldots, 1)'$ on $n \times 1$-vektori. Havaintomatriisi FIRMAT on 10×3-matriisi

$$\underset{10\times 3}{\mathbf{Y}} = \begin{pmatrix} 26.7 & 3.3 & 15.8 \\ 38.4 & 2.4 & 19.5 \\ 19.2 & 1.7 & 8.4 \\ 20.6 & 1.0 & 8.2 \\ 18.9 & 0.9 & 9.4 \\ 14.8 & 1.0 & 7.6 \\ 19.0 & 2.7 & 12.6 \\ 14.2 & 0.8 & 7.3 \\ 13.7 & 1.1 & 5.9 \\ 7.7 & 0.2 & 2.9 \end{pmatrix},$$

jonka 3. havainto on esimerkiksi $\mathbf{y}_3 = (19.2, 1.7, 8.4)'$.

Esimerkki 2.1 Oheisessa R-esimerkissä luetaan Taulukon 2.1 havaintomatriisi tiedostosta `firmat.dat` ja tallennetaan se työtilaan nimellä "firmat". Sen jälkeen tehdään aineistosta numeerinen yhteenveto.

```
firmat<-read.table("C:\\Kurssit\\Mmm\\mmm03\\Datat\\firmat.dat",
header = T)

firmat
         varat ntulot parvo
```

```
G.M.      26.7    3.3  15.8
Exxon     38.4    2.4  19.5
Ford      19.2    1.7   8.4
Mobil     20.6    1.0   8.2
Texaco    18.9    0.9   9.4
Std.Oil   14.8    1.0   7.6
IBM       19.0    2.7  12.6
Gulf      14.2    0.8   7.3
G.E.      13.7    1.1   5.9
Chrysler   7.7    0.2   2.9

summary(firmat)
     varat           ntulot          parvo
 Min.   : 7.70   Min.   :0.200   Min.   : 2.900
 1st Qu.:14.35   1st Qu.:0.925   1st Qu.: 7.375
 Median :18.95   Median :1.050   Median : 8.300
 Mean   :19.32   Mean   :1.510   Mean   : 9.760
 3rd Qu.:20.25   3rd Qu.:2.225   3rd Qu.:11.800
 Max.   :38.40   Max.   :3.300   Max.   :19.500
```

Huomattakoon, että apply-funktiolla voidaan havaintoaineistosta laskea kätevästi erilaisia tunnuslukuja. Seuraavassa lasketaan esimerkiksi muuttujien keskiarvot ja minimit.

```
col.means <- apply(firmat, 2, mean)
col.means
 varat ntulot  parvo
 19.32   1.51   9.76

col.min<- apply(firmat, 2, min)
col.min
 varat ntulot  parvo
   7.7    0.2    2.9
```

Seuraavassa lasketaan muuttujien keskiarvot matriisioperaationa.

```
yks<-rep(1,10)
yks
[1] 1 1 1 1 1 1 1 1 1 1
yks<-as.matrix(yks)
dim(yks)
#[1] 10  1
Y<-as.matrix(firmat)
dim(Y)
```

```
[1] 10  3
n<-dim(Y)[1]
t(Y)%*%yks/n

         [,1]
  varat  19.32
  ntulot  1.51
  parvo   9.76
```

□

2.2 Kovarianssi- ja korrelaatiomatriisimatriisi

Otoskovarianssimatriisi $\mathbf{S} = (s_{ij})$ on p :n muuttujan variansseista ja kovariansseista muodostettu matriisi

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}.$$

Muuttujan y_i otosvarianssi

$$\begin{aligned} s_{ii} = s_i^2 &= \frac{1}{n-1}[(y_{1i} - \bar{y}_i)^2 + (y_{2i} - \bar{y}_i)^2 + \ldots + (y_{ni} - \bar{y}_i)^2] \\ &= \frac{1}{n-1}\sum_{k=1}^{n}(y_{ki} - \bar{y}_i)^2 \\ &= \frac{1}{n-1}(\sum_{k=1}^{n} y_{ki}^2 - n\bar{y}_i^2), \end{aligned}$$

missä $\bar{y}_i$ on i. muuttujan otoskeskiarvo. Huomaa, että varianssi on lausuttu otoshajonnan eli standardipoikkeaman

$$s_i = \sqrt{s_{ii}}$$

neliönä. Muuttujien y_i ja y_j välinen otoskovarianssi on

$$\begin{aligned} s_{ij} &= \frac{1}{n-1}\sum_{k=1}^{n}(y_{ki} - \bar{y}_i)(y_{kj} - \bar{y}_j) \\ &= \frac{1}{n-1}[(y_{1i} - \bar{y}_i)(y_{1j} - \bar{y}_j) + \ldots + (y_{ni} - \bar{y}_i)(y_{nj} - \bar{y}_j)] \\ &= \frac{1}{n-1}(\sum_{k=1}^{n} y_{ki}y_{kj} - n\bar{y}_i\bar{y}_j). \end{aligned}$$

Havaintovektoreiden avulla lausuttuna

$$\begin{aligned}\mathbf{S} &= \tfrac{1}{n-1}\sum_{k=1}^{n}(\mathbf{y}_k - \bar{\mathbf{y}})(\mathbf{y}_k - \bar{\mathbf{y}})' \\ &= \tfrac{1}{n-1}\sum_{k=1}^{n}(\mathbf{y}_k\mathbf{y}_k' - n\bar{\mathbf{y}}\bar{\mathbf{y}}').\end{aligned}$$

Muuttujien y_i ja y_j välinen korrelaatiokerroin on

$$r_{ij} = \frac{s_{ij}}{s_i s_j}. \tag{2.2.2}$$

Otoskorrelaatiomatriisi on muodoltaan vastaava kuin otoskovarianssimatriisi, mutta nyt kovarianssien paikalla ovat korrelaatiot:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}.$$

Kuten yhtälöstä (2.2.2) nähdään, korrelaatiomatriisi saadaan laskettua kovarianssimatriisista. Kovarianssimatriisi saadaan annetusta korrelaatiomatriisista, mikäli hajonnat tunnetaan. Määritellään

$$\begin{aligned}\mathbf{D}_s &= \operatorname{diag}(s_1, s_2, \dots, s_p) \\ &= \begin{pmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_p \end{pmatrix}.\end{aligned}$$

Silloin

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

ja

$$\mathbf{S} = \mathbf{D}_s\mathbf{R}\mathbf{D}_s.$$

Esimerkki 2.2 Tarkastellaan nyt lyhyesti kovarienssien ja korrelaatioiden laskemista.

```
Lasketaan ensin kovarianssimatriisi

cov(firmat)
           varat    ntulot     parvo
varat  70.410667 5.8731111 39.065333
ntulot  5.873111 0.9698889  4.114889
parvo  39.065333 4.1148889 24.056000
```

```
ja sitten korrelaatiomatriisi

cor(firmat)
           varat    ntulot     parvo
varat  1.0000000 0.7107028 0.9492063
ntulot 0.7107028 1.0000000 0.8518937
parvo  0.9492063 0.8518937 1.0000000

Yksittäiset kovarianssit tai korrelaatiot saadaan viittaamalla
kyseisten muuttujien nimiin, kuten esimerkiksi

cov(varat,ntulot); cor(ntulot,parvo)
[1] 5.873111
[1] 0.8518937.
```

Korrelaatiomatriisista (tai kovarianssimatriisista) saadaan yksittäinen alkio käyttämällä tavanomaista matriisien indeksointia. Muuttujien `varat` ja `ntulot` välinen korrelaatio tulostetaan seuraavasti:

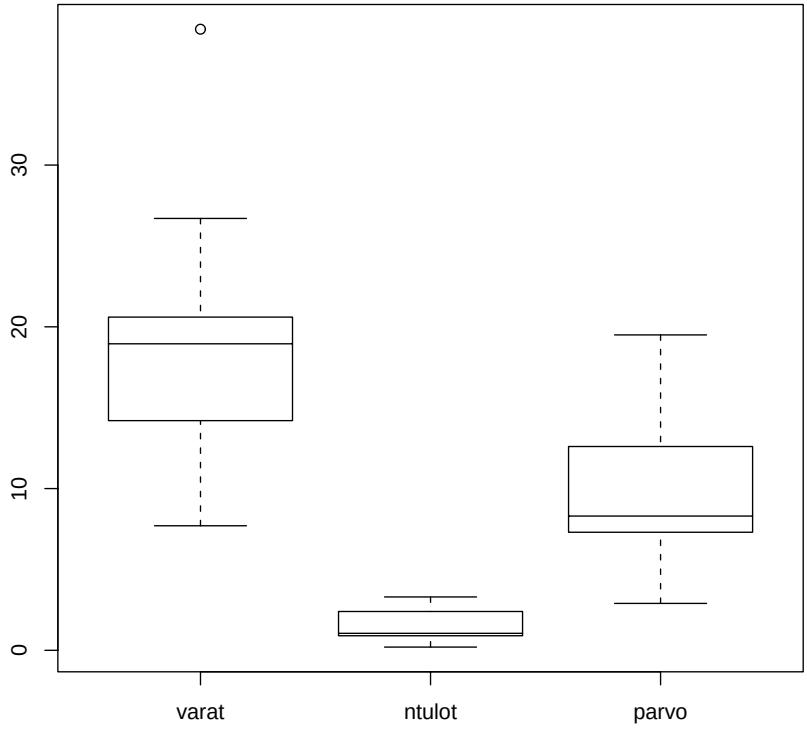
```
cor(firmat)[1,2]
[1] 0.7107028.

Lasketaan sitten varianssit ja hajonnat:
S<-cov(firmat)
diag(S)
     varat     ntulot      parvo
70.4106667  0.9698889 24.0560000

sqrt(diag(S))
    varat    ntulot     parvo
8.3911064 0.9848294 4.9046916
```

□

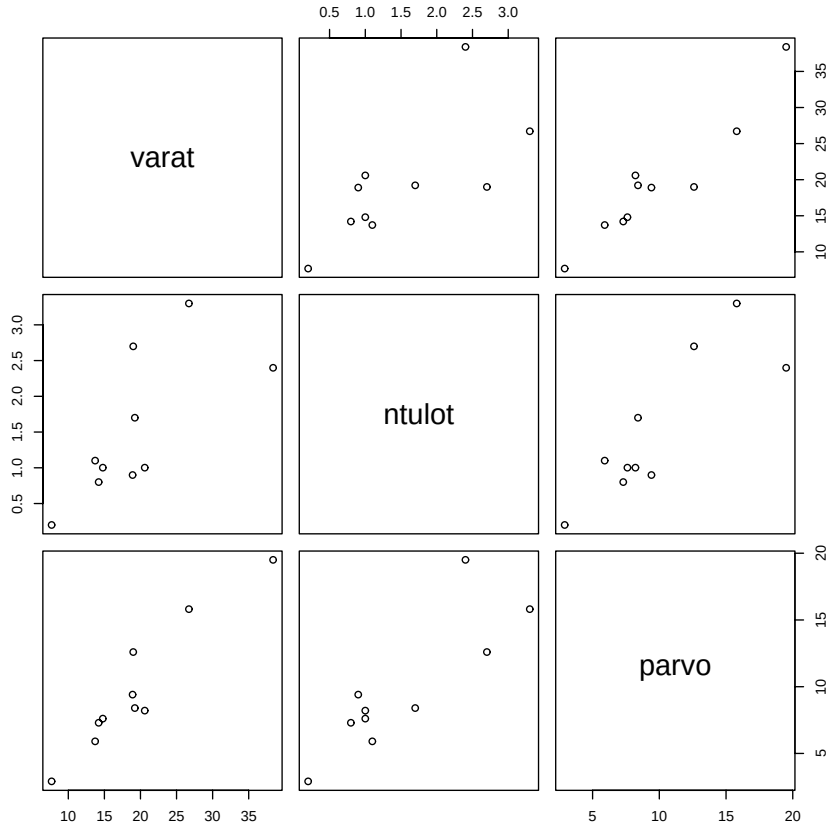
2.3 Graafinen esitys

Seuraavassa tarkastellaan joitain monimuuttujaisen aineiston graafisia esityksiä. Eräs graafinen yhteenveto saadaan komennolla `boxplot(firmat)`, joka tulostaa muuttujien `boxplot`-kuvion.

Muuttujien kaikki parittaiset pisteparvet saadaan komennolla `pairs`. Esimerkiksi `pairs(firmat)` tulostaa pisteparvet



Kuva 2.1: `Firmat`-aineiston muuttujien boxplot-kuviot.



Kuva 2.2: `Firmat`-aineiston muuttujien väliset korrelaatiodiagrammit.

2.4 Muuttujaryhmien keskiarvovektorit ja kovarianssimatriisit

Usein tutkimuksissa on kaksi tai useampia muuttujaryhmiä, joiden välisiä riippuvuuksia halutaan tutkia. Esimerkiksi aineistossa HINNATR on annettu joidenkin elintarvikkeiden keskihinnat 25:ssä Yhdysvaltain kaupungissa. Elintarvikkeista kaksi on hedelmiä (appelsiini, tomaatti) ja muut ovat leipä, hampurilainen ja maito. Merkitään näitä muuttujia y_1=leipä, y_2=hampurilainen, y_3=maito sekä x_1=appelsiini ja x_2=tomaatti. Muuttuja on siis ositettu kahteen osavektoriin $\mathbf{y}$ ja $\mathbf{x}$ ja

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ x_1 \\ x_2 \end{pmatrix}.$$

Yleisessä tapauksessa merkitään i. ositettua havaintovektoria

$$\begin{pmatrix} \mathbf{y}_i \\ \mathbf{x}_i \end{pmatrix} = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ip} \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{iq} \end{pmatrix}, \quad i = 1, 2, \ldots, n,$$

kun muuttujia on yhteensä $p + q$ kappaletta. Esimerkissä $p = 3$ ja $q = 2$.

Otoskeskiarvo ja otoskovarianssimatriisi ovat nyt muotoa

$$\begin{pmatrix} \bar{\mathbf{y}} \\ \bar{\mathbf{x}} \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_p \\ \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_q \end{pmatrix}$$

ja

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{pmatrix},$$

missä $\mathbf{S}_{yy}$ on $p \times p$-matriisi, $\mathbf{S}_{yx}$ on $p \times q$, $\mathbf{S}_{xy}$ on $q \times p$ ja $\mathbf{S}_{xx}$ on $q \times q$. Koska $\mathbf{S}$ on symmetrinen, niin

$$\mathbf{S}_{xy} = \mathbf{S}'_{yx}.$$

Korrelaatiomatriisi $\mathbf{R}$ voidaan tietysti osittaa vastaavalla tavalla. Silloin

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{yy} & \mathbf{R}_{yx} \\ \mathbf{R}_{xy} & \mathbf{R}_{xx} \end{pmatrix},$$

missä osamatriisien $\mathbf{R}_{yy}, \mathbf{R}_{yx}, \mathbf{R}_{xy}$ ja $\mathbf{R}_{xx}$ dimensiot ovat samat kuin vastaavien kovarianssimatriisien.

Jos muuttujista muodostetaan k osajoukkoa, niin ositettuna havaintovektori $\mathbf{y}$ on

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_k \end{pmatrix}.$$

Vektorissa $\mathbf{y}_1$ on p_1 muuttujaa, $\mathbf{y}_2$:ssa on p_2 muuttujaa ja $\mathbf{y}_k$:ssa on p_k muuttujaa siten, että $p_1 + p_2 + \ldots + p_k = p$. Otoskeskiarvovektori ja otoskovarianssimatriisi ovat

$$\mathbf{y} = \begin{pmatrix} \bar{\mathbf{y}}_1 \\ \bar{\mathbf{y}}_2 \\ \vdots \\ \bar{\mathbf{y}}_k \end{pmatrix}$$

ja

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & \ldots & \mathbf{S}_{1k} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \ldots & \mathbf{S}_{2k} \\ \vdots & \vdots & & \vdots \\ \mathbf{S}_{k1} & \mathbf{S}_{k2} & \ldots & \mathbf{S}_{kk} \end{pmatrix}.$$

Esimerkiksi osamatriisi $\mathbf{S}_{1k}$ on osavektoreiden $\mathbf{y}_1$ ja $\mathbf{y}_k$ välinen kovarianssimatriisi.

Esimerkki 2.3 Lasketaan `Hinnatr`-aineistosta muuttujaryhmien *sisäiset* ja *väliset* korrelaatiot. Matriisit $\mathbf{R}_{yy}, \mathbf{R}_{xx}$ ja $\mathbf{R}_{yx}$ saadaan komennoilla `cor(hinnatr[,1:3])`, `cor(hinnatr[,4:5])` ja `cor(hinnatr[,1:3],hinnatr[,1:3])`. Komennoissa havaintomatriisia (R:n `data frame`) on käsitelty matriisina. Oheisessissa laskelmissa on ensin tulostettu kymmenen ensimmäisen kaupungin hinnat ja laskettu sitten keskiarvovektorit $\bar{\mathbf{y}}'$ ja $\bar{\mathbf{x}}'$ sekä matriisit $\mathbf{R}_{yy}, \mathbf{R}_{xx}$ ja $\mathbf{R}_{yx}$.

```
hinnatr[1:10,]
           leipa hampuril   voi appels tomaatti
Anchorage   70.9    135.6 155.0   63.9    100.1
Atlanta     36.4    111.5 144.3   53.9     95.9
Baltimore   28.9    108.8 151.0   47.5    104.5
Boston      43.2    119.3 142.0   41.1     96.5
Buffalo     34.5    109.9 124.8   35.6     75.9
Chicago     37.1    107.5 145.4   65.1     94.2
Cincinnati  37.1    118.1 149.6   45.6     90.8
```

```
Cleveland   38.5    107.7 142.7   50.3     83.2
Dallas      35.5    116.8 142.5   62.4     90.7
Detroit     40.8    108.8 140.1   39.7     96.1

apply(hinnatr[,1:3], 2, mean)

    leipa  hampuril       voi
 38.44167 112.24583 144.21292

apply(hinnatr[,4:5], 2, mean)

  appels tomaatti
51.73750 89.75833

cor(hinnatr[,1:3])
             leipa  hampuril       voi
leipa    1.0000000 0.6490532 0.3301770
hampuril 0.6490532 1.0000000 0.2447778
voi      0.3301770 0.2447778 1.0000000

cor(hinnatr[,4:5])
            appels  tomaatti
appels   1.0000000 0.1333844
tomaatti 0.1333844 1.0000000

cor(hinnatr[,1:3],hinnatr[,4:5])
            appels  tomaatti
leipa    0.3187031 0.3620681
hampuril 0.1908956 0.5557993
voi      0.2351424 0.4361291
```

Matriisimerkinnöin tämä ositettu korrelaatiomatriisi on muotoa

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{yy} & \mathbf{R}_{yx} \\ \mathbf{R}_{xy} & \mathbf{R}_{xx} \end{pmatrix} = \left(\begin{array}{ccc|cc} 1.0000 & 0.6491 & 0.3302 & 0.3187 & 0.3621 \\ 0.6491 & 1.0000 & 0.2448 & 0.1909 & 0.5558 \\ 0.3302 & 0.2448 & 1.0000 & 0.2351 & 0.4361 \\ \hline 0.3187 & 0.1909 & 0.2351 & 1.0000 & 0.1334 \\ 0.3621 & 0.5558 & 0.4361 & 0.1334 & 1.0000 \end{array}\right).$$

□

2.5 Muuttujien lineaariset yhdisteet

Monimuuttujamenetelmissä tarkastellaan usein muuttujien $y_1, y_2, \ldots, y_p$ lineaarisia yhdisteitä. Tässä pykälässä esitetään lineaaristen yhdisteiden keskiarvoja, variansseja ja kovariansseja koskevat perustulokset, jotka ovat jatkossa varsin käyttökelpoisia.

Olkoot $a_1, a_2, \ldots, a_p$ vakiota ja tarkastellaan muuttujien $y_1, y_2, \ldots, y_p$ lineaarista yhdistettä

$$z = a_1 y_1 + a_2 y_2 + \ldots + a_p y_p = \mathbf{a}'\mathbf{y},$$

missä $\mathbf{a} = (a_1, a_2, \ldots, a_p)'$ on kerroinvektori ja $\mathbf{y} = (y_1, y_2, \ldots y_p)'$ on p-ulotteinen muuttujavektori. Tämä muuttujien $y_1, y_2, \ldots, y_p$ lineaarinen yhdiste määrittelee uuden muuttujan z. Kertomalla havaintovektori $\mathbf{y}_i$ kerroinvektorilla $\mathbf{a}$ saadaan muuttujan z i. havaintoarvo

$$\begin{aligned} z_i &= a_1 y_{i1} + a_2 y_{i2} + \ldots + a_p y_{ip} \\ &= \mathbf{a}'\mathbf{y}_i, \quad i = 1, 2, \ldots, n. \end{aligned}$$

Muutujan z keskiarvo on arvojen $z_1 = \mathbf{a}'\mathbf{y}_1, z_2 = \mathbf{a}'\mathbf{y}_2, \ldots, z_n = \mathbf{a}'\mathbf{y}_n$ keskiarvo tai keskiarvojen $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_p$ lineaarinen yhdiste eli

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \mathbf{a}'\bar{\mathbf{y}}, \tag{2.5.3}$$

missä $\bar{\mathbf{y}}$ on havaintovektoreiden $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ otoskeskiarvo. Vastaava yhden muuttujan tulos on $\bar{z} = a\bar{y}$, missä $z_i = ay_i,\ i = 1, 2, \ldots, n$. Havaintomatriisin $\mathbf{Y}$ avulla lausuttuna lineaarisen yhdisteen muodostama uusi muuttuja on

$$\mathbf{z} = \mathbf{Y}\mathbf{a},$$

missä

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} = \begin{pmatrix} \mathbf{a}'\mathbf{y}_1 \\ \mathbf{a}'\mathbf{y}_2 \\ \vdots \\ \mathbf{a}'\mathbf{y}_n \end{pmatrix} \quad \text{ja} \quad \mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \ldots & y_{1p} \\ y_{21} & y_{22} & \ldots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \ldots & y_{np} \end{pmatrix}.$$

Muutujan z otosvarianssi saadaan laskemalla z :n arvojen $z_1 = \mathbf{a}'\mathbf{y}_1, z_2 = \mathbf{a}'\mathbf{y}_2, \ldots, z_n = \mathbf{a}'\mathbf{y}_n$ varianssi tai käyttämällä hyväksi havaintovektoreiden $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ otoskovarianssimatriisia $\mathbf{S}$:

$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \mathbf{a}'\mathbf{S}\mathbf{a}. \tag{2.5.4}$$

Yhden muuttujan y tapauksessa $s_z^2 = a^2 s^2$, missä s^2 on y :n varianssi ja $z_i = ay_i,\ i = 1, 2, \ldots, n$.

Neliösummana varianssi s_z^2 ei voi olla negatiivinen. Siksi $s_z^2 \geq 0$ ja vastaavasti $\mathbf{a}'\mathbf{S}\mathbf{a} \geq 0$ kaikilla vektorin $\mathbf{a}$ arvoilla. Tästä seuraa, että $\mathbf{S}$ on epänegatiivisesti definiitti. Jos muuttujat ovat jatkuvia ja lineaarisesti riippumattomia sekä $n - 1 \geq p$, niin $\mathbf{S}$ on positiivisesti definiitti (todennäköisyydellä 1).

Määritellään nyt lineaariset yhdisteet $z_1 = \mathbf{a}_1'\mathbf{y} = a_{11}y_1 + a_{12}y_2 + \ldots + a_{1p}y_p$ ja $z_2 = \mathbf{a}_2'\mathbf{y} = a_{21}y_1 + a_{22}y_2 + \ldots + a_{2p}y_p$, missä $\mathbf{a}_1 = (a_{11}, a_{12}, \ldots, a_{1p})'$ ja $\mathbf{a}_2 = (a_{21}, a_{22}, \ldots, a_{2p})'$ ovat kaksi eri vakiovektoria. Muuttujien z_1 ja z_2 otoskovarianssi on

$$s_{z_1 z_2} = \mathbf{a}_1'\mathbf{S}\mathbf{a}_2. \tag{2.5.5}$$

Tulosten (2.5.4) ja (2.5.5) nojalla z_1 :n ja z_2 :n välinen otoskorrelaatio on

$$r_{z_1 z_2} = \frac{s_{z_1 z_2}}{\sqrt{s_{z_1}^2 s_{z_2}^2}} = \frac{\mathbf{a}_1'\mathbf{S}\mathbf{a}_2}{\sqrt{(\mathbf{a}_1'\mathbf{S}\mathbf{a}_1)(\mathbf{a}_2'\mathbf{S}\mathbf{a}_2)}}. \tag{2.5.6}$$

Esitetään nyt lineaariset yhdisteet z_1 ja z_2 $2 \times p-$matriisin

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \end{pmatrix}$$

avulla $2 \times 1-$vektorina:

$$\begin{aligned} \mathbf{z} &= \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1'\mathbf{y} \\ \mathbf{a}_2'\mathbf{y} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \end{pmatrix} \mathbf{y} = \mathbf{A}\mathbf{y}. \end{aligned}$$

Kahden muuttujan havaintoarvot $\mathbf{z}_i = \mathbf{A}\mathbf{y}_i,\ i = 1, 2, \ldots, n$ saadaan siis lineaarisena muunnoksena $p-$ulotteisista havainnoista $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$.

Muuttujan $\mathbf{z} = (z_1, z_2)$ otoskeskiarvo voidaan laskea $\bar{\mathbf{y}}$:n avulla:

$$\begin{aligned} \bar{\mathbf{z}} &= \begin{pmatrix} \bar{z}_1 \\ \bar{z}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1'\bar{\mathbf{y}} \\ \mathbf{a}_2'\bar{\mathbf{y}} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \end{pmatrix} \bar{\mathbf{y}} = \mathbf{A}\bar{\mathbf{y}}. \end{aligned}$$

Vastaavasti $\mathbf{z}$:n otoskovarianssimatriisi on

$$\begin{aligned} \mathbf{S}_z &= \begin{pmatrix} s_{z_1}^2 & s_{z_1 z_2} \\ s_{z_2 z_1} & s_{z_2}^2 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{a}_1'\mathbf{S}\mathbf{a}_1 & \mathbf{a}_1'\mathbf{S}\mathbf{a}_2 \\ \mathbf{a}_2'\mathbf{S}\mathbf{a}_1 & \mathbf{a}_2'\mathbf{S}\mathbf{a}_2 \end{pmatrix} = \mathbf{A}\mathbf{S}\mathbf{A}'. \end{aligned}$$

Tulokset

$$\begin{aligned} \bar{\mathbf{z}} &= \mathbf{A}\bar{\mathbf{y}} \\ \mathbf{S}_z &= \mathbf{A}\mathbf{S}\mathbf{A}' \end{aligned}$$

ovat jatkossa erittäin hyödyllisiä. Ne pitävät paikkansa yleisesti k:lle lineaariselle yhdisteelle $z_1 = \mathbf{a}_1'\mathbf{y}, z_2 = \mathbf{a}_2'\mathbf{y}, \ldots, z_k = \mathbf{a}_k'\mathbf{y}$, jolloin $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2 \ldots \mathbf{a}_k)'$ on $k \times p$-matriisi. Vastaava tulos pätee hieman yleisemmälle lineaariselle muunnokselle

$$\mathbf{z}_i = \mathbf{A}\mathbf{y}_i + \mathbf{b}; \quad i = 1, 2, \ldots, n, \tag{2.5.7}$$

missä $\mathbf{b} = (b_1, b_2, \ldots, b_p)'$ on vakiovektori. Otoskeskiarvovektori ja kovarianssimatriisi ovat

$$\bar{\mathbf{z}} = \mathbf{A}\bar{\mathbf{y}} + \mathbf{b} \tag{2.5.8}$$

$$\mathbf{S}_z = \mathbf{A}\mathbf{S}\mathbf{A}'. \tag{2.5.9}$$

Esimerkki 2.4 Olkoon $p = 2$. Silloin $\mathbf{y} = (y_1, y_2)'$, $z = a_1y_1 + a_2y_2$ ja

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix},$$

missä $s_{12} = s_{21}$. Laskemalla todetaan, että

$$\begin{aligned} s_y^2 &= (a_1, a_2) \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \\ &= (a_1s_{11} + a_2s_{12}, a_1s_{12} + a_2s_{22}) \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \\ &= a_1^2s_{11} + 2a_1a_2s_{12} + a_2^2s_{22}. \end{aligned}$$

Jos esimerkiksi kovarianssimatriisi

$$\mathbf{S} = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$$

tunnetaan, niin muuttujan $z = 5y_1 + 10y_2$ varianssi

$$s_z^2 = 5^2 \cdot 2 + 2 \cdot 5 \cdot 10 \cdot 1 + 10^2 \cdot 3 = 255$$

voidaan laskea tuntematta lainkaan havaintomatriisia $\mathbf{Y}$. □

Esimerkki 2.5 Tarkastellaan vielä `Hinnatr`-aineistoa. Muodostetaan elintarvikkeista kaksi erilaista koria, joissa toisessa on suhteellisesti enemmän leipätuotteita ja toisessa taas suhteellisesti enemmän hedelmiä. Muodostetaan uudet muuttujat z_1 ja z_2 siten, että z_1 on 1. korin hinta ja z_2 on 2. korin hinta. Vektori $\mathbf{a}_1$ sisältää eri tuotteiden suhteelliset osuudet 1. korissa ja $\mathbf{a}_2$ suhteelliset osuudet 2. korissa.

```
a1<-c(1/5,2/5,1/10,1/10,1/5)
a2<-c(1/5,1/10,2/10,3/10,1/5)

Muutetaan hinnatr sekä a1 ja a1 matriisiksi.
Y<-as.matrix(hinnatr)
```

```
a1<-as.matrix(a1)
a2<-as.matrix(a2)

Muodostetaan muuttujat z1 ja z2.
z1<-Y%*%a1; z2<-Y%*%a2

Sama voidaan tehdä yhdellä matriisioperaatiolla, kunhan ensin on muodostetaan
tarvittava kerroinmatriisi A.

A<-rbind(t(a1),t(a2))
Z<-Y%*%t(A)
```

Seuraavassa muodostetaan uusi havaintomatriisi hinnat, jossa havaintomatriisin hinnatr muuttujien lisäksi on uudet muuttujat z_1 ja z_2.

```
hinnat<-transform(hinnatr,z1=Y%*%a1,z2=Y%*%a2)
```

Muuttujien z_1 ja z_2 kovarianssimatriisi havaintomatriisista laskettuna on

```
cov(hinnat[,6:7])
         z1       z2
z1 55.12915 38.53941
z2 38.53941 34.69713
```

Koska z_1 ja z_2 ovat havaintomatriisin hinnatr muuttujien tunnettuja lineaarikombinaatioita, niin niiden kovarianssimatriisi on

```
A%*%cov(hinnatr)%*%t(A)

           [,1]     [,2]
  [1,] 55.12915 38.53941
  [2,] 38.53941 34.69713
```

□

2.6 Kokonaisvaihtelun mittoja

Kovarianssimatriisi sisältää muuttujien variannssit ja kovarianssit, joten sen avulla saa kokonaiskuvan aineiston muuttujien vaihtelusta. Usein tätä vaihtelua halutaan kuvata yhdellä tunnusluvulla. Tavanomaisimmat monimuuttujaisen aineiston kokonaisvaihtelun mitat ovat *yleistetty varianssi* ja *kokonaisvarianssi*. Yleistetty varianssi on kovarianssimatriisin $\mathbf{S}$ determinantti ja kokonaisvarianssi on kovarianssimatriisin jälki:

$$\text{Yleistetty otosvarianssi} = |\mathbf{S}|$$

ja

$$\text{Otoksen kokonaisvarianssi} = s_1^2 + s_2^2 + \ldots + s_p^2 = \text{tr}(\mathbf{S}).$$

Olkoot $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ kovarianssimatriisin $\mathbf{S}$ ominaisarvot. Koska $\mathbf{S}$ on epänegatiivisesti definiitti, niin sen ominasiarvot ovat epänegatiiviset. Matriisin determinatti on sen ominaisarvojen tulo, joten

$$|\mathbf{S}| = \lambda_1 \lambda_2 \ldots \lambda_p.$$

Yleistetyllä varianssilla on geometrinen tulkinta. Muodostetaan $\mathbf{S}$:n avulla p-ulotteinen hyperellipsoidi

$$(\mathbf{y} - \bar{\mathbf{y}})' \mathbf{S}^{-1} (\mathbf{y} - \bar{\mathbf{y}}) = k, \tag{2.6.10}$$

jonka keskipiste on otoskeskiarvo $\bar{\mathbf{y}}$. Ellipsoidin pääakselien puolikkaiden pituudet ovat $\sqrt{k\lambda_1}, \sqrt{k\lambda_2}, \ldots, \sqrt{k\lambda_p}$ ja pääakselit ovat ominaisarvoja vastaavien ominaisvektorien suuntaiset. Ellipsoidin tilavuus on $c(p,k)\sqrt{|\mathbf{S}|} = c(p,k)\sqrt{\lambda_1 \lambda_2 \ldots \lambda_p}$, missä vakio $c(p,k)$ riippuu vain muuttujien lukumääristä p ja annetusta vakiosta k. Ellipsoidin tilavuus riippuu siis yleistettystä varianssista $|\mathbf{S}|$. Jos $\lambda_p = 0$, niin $|\mathbf{S}| = 0$. Silloin joidenkin muuttujien välillä on lineaarinen riippuvuus, joten havaintomatriisin aste onkin vain $p-1$. Tässä tapauksessa hyperellipsoidi (2.6.10) on kokonaan jossain $p-1$-ulotteisessa aliavaruudessa ja ellipsoidin tilavuus on "litistynyt" nollaksi. Eräs tapa "ratkaista" tämä pulma on poistaa jokin tai joitain muuttujia aineistosta siten, että muuttujien välillä ei ole enää lineaarisia riipuvuuksia. Kovarianssimatriisin ja korrelaatiomatriisin välisen yhteyden nojalla

$$|\mathbf{S}| = s_1^2 s_2^2 \ldots s_p^2 |\mathbf{R}|.$$

Mitä pienempi siis $|\mathbf{R}|$ on, sitä pienempi on kokonaisvarianssia (mikäli varianssit pysyvät muuttumattomina).

Kokonaisvarianssi $\text{tr}(\mathbf{S}) = s_1^2 + s_2^2 + \ldots + s_p^2$ on siis yksinkertaisesti yksittäisten muuttujien varianssien summa. Tämä vaihtelun mitta ei ota muuttujien välistä riippuvuutta lainkaan huomioon. Edellä mainittujen mittojen suuret arvot tarkoittavat havaintojen $\mathbf{y}_i$ suurta vaihtelua keskiarvon $\bar{\mathbf{y}}$ ympärillä.

Esimerkki 2.6 Tarkastellaan nyt havaintoaineiston `Hinnatr` muuttujien kokonaisvaihtelua.

```
S<-cov(hinnatr)
```

Muuttujien varianssit ovat

```
diag(S)
    leipa  hampuril       voi    appels  tomaatti
 69.99906 135.33042  85.13557  69.87288  54.74341
```

```
ja kokonaisvarianssi on niiden summa
sum(diag(S))
[1] 415.0813.

Yleistetty varianssi on
det(S)
[1] 822907260,

varianssien tulo
prod(diag(S))
[1] 3084885998

ja korrelaatiomatriisin determinantti on
det(cor(hinnatr))
[1] 0.2667545

Kovarianssimatriisin ja korrelaatiomatriisin välisen yhteyden nojalla
voidaan todeta, että yleistetty varianssi det(S) on edellisten tulo, eli

prod(diag(S))*det(cor(hinnatr))
[1] 822907260

Muodostetaan nyt hinnatr-aineistoa muuttujien lineaarinen yhdiste,
jonka kerroinvaktori on
a1<-c(1/5,2/5,1/10,1/10,1/5)
Y<-as.matrix(hinnatr)
a1<-as.matrix(a1)

Lisätään tämä muuttuja havaintoaineistoon ja lasketaan yleistetty
varianssi.

hinnat<-transform(hinnatr,z1=Y%*%a1)
det(cov(hinnat))
[1] -1.230630e-05
round(det(cov(hinnat)),4)
[1] 0
```

Yleistetty varianssi on 0 (laskennassa päästy neljän desimaalin tarkkuuteen), koska muuttuja z1 on muiden muuttujien lineaarinen yhdiste. Havaintomatriisi ja kovarinassimatriisi eivät ole siis täysasteisia.

Sen sijaan hinnat-aineiston kokonaisvarianssi

```
sum(diag(cov(hinnat)))
[1] 470.2105
```

```
on suurempi kuin hinnatr-aineiston kokonaisvarianssi
sum(diag(cov(hinnatr)))
[1] 415.0813
```

Näiden kahden mitan käyttäytyminen on siis keskenään ristiriitaista, mutta tiedämme siihen syyn. □

Annamme yleistetylle varianssille vielä toisenkin geometrisen tulkinnan. Olkoot $\mathbf{y}_1$ ja $\mathbf{y}_2$ annetut keskistetyt havaintovektorit ja olkoon niiden välinen kulma α. Skaalataan vektorit jakamalla ne luvulla $\sqrt{n-1}$. Tarkastellaan suunnikasta, jonka kyljet ovat $\frac{\mathbf{y}_1}{\sqrt{n-1}}$ ja $\frac{\mathbf{y}_1}{\sqrt{n-1}}$. Suunnikaan ala A on

$$A = (\frac{\|\mathbf{y}_1\|\|\mathbf{y}_2\|}{n-1} sin\,\alpha)^2. \tag{2.6.11}$$

Tämä ala on samalla muuttujien $\mathbf{y}_1$ ja $\mathbf{y}_2$ kovarianssimatriisin determinatti. Yleisessä p:n muuttujan tapauksessa yleistetty varianssi on vastaavalla tavalla p-ulotteisessa avaruudessa määritellyn suuntaissärmiön tilavuus.

2.7 Havaintojen esittäminen uusien akselien suhteen

Monimuuttujaisissa tekniikoissa havainnot usein esitetään uuden koordinaatiston suhteen. Oletetaan, että tarkasteltavat koordinaatistot ovat ortonormaalisia. Olkoot $\mathbf{e}_1 = (1,0)'$ ja $\mathbf{e}_2 = (0,1)'$ luonnolliset kantavektorit, jotka esittävät vastaavia akseleita x_1 ja x_2. Vektorit $\mathbf{e}_1$ ja $\mathbf{e}_2$ ovat ortonormaaliset, sillä $\|\mathbf{e}_1\| = \|\mathbf{e}_2\| = 1$ ja $\mathbf{e}_1'\mathbf{e}_2 = 0$. Esittäköön vektori $\mathbf{a} = (a_1, a_2)'$ (havainto)pistettä A, joten

$$\mathbf{a} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2.$$

Muodostetaan nyt uudet koordinaattiakselit kiertämällä akseleita x_1 ja x_2 vastapäivään kulman θ verran. Merkitään uusia akseleita x_1^* ja x_2^*. Olkoon $\mathbf{a}^* = (a_1^*, a_2^*)'$ pisteen A koordinaattiesitys akselistossa x_1^* ja x_2^* (piste A pysyy paikallaan). On helppo osoittaa, että

$$\mathbf{a}^* = \begin{pmatrix} cos\,\theta & sin\,\theta \\ -sin\,\theta & cos\,\theta \end{pmatrix} \mathbf{a}. \tag{2.7.12}$$

Yleisesti, koordinaatiston kierto saadaan aikaan kertomalla koordinaattivektori ortogonaalisella matriisilla $\mathbf{P}$. Matriisi $\mathbf{P}$ on ortogonaalinen, jos $\mathbf{P}'\mathbf{P} = \mathbf{P}\mathbf{P}' = \mathbf{I}$.

Harjoituksia

1. Laske FIRMAT-aineistosta (Yhdysvaltain 10 suurinta yritystä 1978) keskiarvovektori sekä kovarianssi- ja korrelaatiomatriisit. Muodosta korrelaatiodiagrammit.

2. Totea FIRMAT-aineistosta saatuja lukuarvoja käyttäen, että esimerkiksi
   (a) $r_{12} = \frac{s_{12}}{s_1 s_2}$ ja $r_{23} = \frac{s_{23}}{s_2 s_3}$.
   (b) Tee sijoitukset (mittayksikön muunnokset) $y_1 := 2y_1,\ y_2 := 3y_2$ ja $y_3 := 5y_3$, kun y_1=varat, y_2=tulot ja y_3=arvo. Mitä ovat muunnettujen muuttujien hajonnat, kovarianssit ja korrelaatiot?
   (c) Laske s_z, kun $z = 2y_1 + 3y_2 + 5y_3$.

3. Olkoon havaintomatriisi
$$\mathbf{Y} = \begin{pmatrix} -1 & 3 & -2 \\ 2 & 4 & 2 \\ 5 & 2 & 3 \end{pmatrix}.$$
   (a) Keskistä ja
   (b) standardoi havainnot.
   (c) Laske $\mathbf{S}$:n determinantti $|\mathbf{S}|$ (yleistetty varianssi).

4. Merkitään
$$\mathbf{D} = \begin{pmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_3 \end{pmatrix},$$
   missä s_i on i. $(i = 1, 2, 3)$ muuttujan hajonta.
   (a) Näytä, että $\mathbf{S} = \mathbf{DRD}$ ja $\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}$, kun $\mathbf{R}$ on korrelaatiomatriisi ja
$$\mathbf{D}^{-1} = \begin{pmatrix} s_1^{-1} & 0 & 0 \\ 0 & s_2^{-1} & 0 \\ 0 & 0 & s_3^{-1} \end{pmatrix}$$
   on $\mathbf{D}$:n käänteismatriisi.
   (b) Lausu $\mathbf{S}$:n determinantti hajontojen ja $\mathbf{R}$:n determinantin avulla.
   (c) Tarkista tulos 3. tehtävän datan avulla.

5. Olkoon $\mathbf{a}' = (a_1, a_2, a_3)$, missä $a_1 = \frac{1}{2}$ ja $a_2 = a_3 = \frac{1}{4}$. Laske muuttujan $z = \mathbf{a}'\mathbf{y} = a_1 y_1 + a_2 y_2 + a_3 y_3$
   (a) keskiarvo ja
   (b) varianssi 3. tehtävän aineistosta.

6. Piirrä 3. tehtävän aineistosta
   (a) pisteparvi $(y_{i1}, y_{i2}),\ i = 1, 2, 3$ sekä
   (b) keskistettyjen ja
   (c) standardoitujen havaintojen pisteparvi koordinaatistoon.

7. Osoita, että

   (a) $\bar{z} = \mathbf{a}'\bar{\mathbf{y}}$ ja

   (b) $s_z^2 = \mathbf{a}'\mathbf{S}\mathbf{a}$, kun $z_i = \mathbf{a}'\mathbf{y}_i, i = 1, 2, \ldots, n$.

8. Määritellään lineaariset yhdisteet

$$\begin{aligned} z_1 &= y_1 + y_2 + y_3 \\ z_2 &= 2y_1 - 3y_2 + 2y_3 \\ z_3 &= -y_1 - 2y_2 - 3y_3. \end{aligned}$$

   Lausu
   (a) $\bar{\mathbf{z}}$ ja $\mathbf{S}_z$ $\mathbf{y}$:n kovarianssimatriisin avulla.
   (b) Lausu $\mathbf{R}_z$ kovarianssimatriisin $\mathbf{S}_z$ avulla.

9. Laske LUU-aineistosta
   (a) $\bar{\mathbf{y}}$, $\mathbf{S}$ ja $\mathbf{R}$ sekä
   (b) $|\mathbf{S}|$ ja $\mathrm{tr}(\mathbf{S})$.

```
read.table(" ... \\tilasto\\mmm\\Datat\\Luu.txt", header = T, skip=3)
```

10. Merkitään POJAT-aineiston ensimmäisten poikien pään mittoja $\mathbf{y} = (y_1, y_2)'$ ja toisten poikien vastaavia mittoja $\mathbf{x} = (x_1, x_2)'$

    (a) Laske keskiarvot ja lausu ne ositettuna vektorina $\begin{pmatrix} \bar{\mathbf{y}} \\ \bar{\mathbf{x}} \end{pmatrix}$.

    (b) Laske kovarianssimatriisi ja lausu se ositettuna

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{pmatrix}.$$

11. Olkoon havaintomatriisi

$$\mathbf{Y} = \begin{pmatrix} -1 & 3 \\ 2 & 4 \\ 5 & 2 \end{pmatrix}.$$

    Osoita laskemalla, että kaava (2.6.11) pitää tässä tapauksessa paikkansa (Osoita: $|\mathbf{S}|$ = suunnikkaan ala). Hahmottele suunnikas.

12. (a) Pisteiden A ja B koordinaatit ortogonaalisten akseleiden x_1 ja x_2 suhteen ovat $A : \ (3, -2)'; B : \ (5, 1)'$. Jos akseleita kierretään 20° vastapäivään, niin mitkä ovat A:n ja B:n koordinaatiti uusien akselien x_1^* ja x_2^* suhteen.

    (b) Pisteen A koordinaatit ortogonaalisten akseleiden x_1 ja x_2 suhteen $(5, 1)'$. Kierretään akseleita myötäpäivään kulman θ verran. Pisteen A koordinaatit uusien kierrettyjen akselien suhteen ovat $(3.69, 3.93)$. Laske θ.

2.8 Aineistot

FIRMAT
HINNATR
LUU
POJAT

Luku 3

Pääkomponenttianalyysi

3.1 Johdanto

Pääkomponenttianalyysissa pyritään löytämään pienehkö määrä alkuperäisten muuttujien lineaarisia yhdisteitä, jotka selittävät mahdollisimman paljon muuttujien kokonaisvaihtelusta. Haluamme esimerkiksi asettaa oppilaat järjestykseen matematiikan, fysiikan ja kielten arvosanojen perusteella. Tavanomainen tapa on käyttää vertailuperusteena oppiaineiden arvosanojen aritmeettista keskiarvoa, joka on eräs muuttujien lineaarinen yhdiste. Selvimmät erot oppilaiden välille saadaan kuitenkin käyttämällä keskiarvon sijasta aineistosta määritettyä ensimmäistä pääkomponenttia.

Usein pääkomponenttianalyysia käytetään muiden analyysien apuna vähentämään tutkittavien muuttujien lukumäärää. Esimerkiksi regressioanalyysissa voi olla (1) liian monta selittävää muuttujaa suhteessa havaintojen lukumäärään tai (2) selittäjät ovat voimakkaasti korreloituneita. Tällaisissa tapauksissa päästään usein luotettavampiin tuloksiin valitsemalla selittäjiksi vähäinen määrä pääkomponentteja.

3.2 Pääkomponenttianalyysin perusteet

Havaintomatriisissa

$$\underset{n\times p}{\mathbf{Y}} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix}$$

on n:n havaintovektorin otos, josta lasketaan $\bar{\mathbf{y}}$ ja $\mathbf{S}$. Jos muuttujat y_1, $y_2, \ldots, y_p$ ovat korreloituneita, havaintojen pisteparvi ei ole koordinaattiakselien suuntainen. Pääkomponenttianalyysissa etsitään pisteparven luonnolliset akselit siten, että origo on pistessaä $\bar{\mathbf{y}}$ ja akselit ovat parven suuntaisia. Keskistetyt havainnot ovat muotoa $\mathbf{y} - \bar{\mathbf{y}}$.

Esimerkiksi aineistossa ASANAT on 30 oppilaan englannin (y_1), ruotsin (y_2), algebran (y_3), geometian (y_4), fysiikan (y_5) ja saksan (y_6) arvosanat. Tavallisesti kokonaissuorituksia vertaillaan arvosanojen aritmeettisen keskiarvon

$$v = \frac{1}{6}\mathbf{1}'\mathbf{y} = \frac{1}{6}y_1 + \frac{1}{6}y_2 + \ldots + \frac{1}{6}y_6$$

perusteella. Aritmeettisessa keskiarvossa painotetaan kaikkia aineita yhtä paljon eli arvosanojen kerroinvektori $\mathbf{a} = \frac{1}{6}(1,1,1,1,1,1)'$. Luovumme nyt tästä rajoitteesta ja tarkastelemme kaikkia painotettuja keskiarvoja.

Tarkastelemme kysymystä: Mikä on se arvosanojen painotettu keskiarvo, jolla on suurin otosvarianssi? Olkoon kerroinvektorin $\mathbf{a}_1$ pituus kiinnitetty ykköseksi ($\mathbf{a}_1'\mathbf{a}_1 = 1$). Ensimmäinen pääkomponentti $z_1 = \mathbf{a}_1'\mathbf{y}$ on sellainen muuttujien $y_1, \ldots, y_p$ lineaarinen yhdiste ($\mathbf{a}_1'\mathbf{a}_1 = 1$), jolla on suurin varianssi. ASANAT-aineistossa

$$z_1 = 0.486y_1 + 0.351y_2 + 0.396y_3 + 0.590y_4 + 0.344y_5 + 0.133y_6,$$

missä kerroinvektori $\mathbf{a}_1' = (0.486, 0.351, 0.396, 0.590, 0.344, 0.133)$ painottaa eri oppiaineita eri tavoin. Koska kerroinvektorin $\mathbf{a}_1$ alkiot ovat positiivisia, voidaan 1. pääkomponentti muuntaa painotetuksi keskiarvoksi jakamalla kertoimet $a_{11}, \ldots, a_{16}$ kertoimien summalla (Painotetussa keskiarvossa kertoimet ovat positiivisia ja niiden summa on 1). Saatu painotettu keskiarvo on

$$w_1 = 0.211y_1 + 0.153y_2 + 0.172y_3 + 0.257y_4 + 0.149y_5 + 0.058y_6.$$

Arvosanojen aritmeettisen keskiarvon otoshajonta $s_{ka} = 0.688$ on selvästi painotetun keskiarvon w_1 otoshajontaa $s_{w_1} = 0.764$ pienempi.

Otoksesta $\mathbf{y}_1$, $\mathbf{y}_2$, ..., $\mathbf{y}_n$ voimme laskea keskiarvon $\bar{\mathbf{y}}$ ja kovarianssimatriisin $\mathbf{S}$. Pääkomponentit lasketaan kovarianssimatriisista. Koska havaintoaineiston siirtäminen eli saman vakiovektorin lisääminen kaikkiin havaintovektoreihin ei vaikuta kovarianssimatriisiin, tarkastellaan keskistettyjä havaintoja $\mathbf{y}_1 - \bar{\mathbf{y}}$, $\mathbf{y}_2 - \bar{\mathbf{y}}$, ..., $\mathbf{y}_n - \bar{\mathbf{y}}$. Pääkomponentit määrittävät keskistetylle aineistolle uudet koordinaattiakselit (pääakselit) ja saadut uudet muuttujat (pääkomponentit) ovat korreloimattomia. Merkintöjen yksinkertaistamiseksi oletetaan tässä luvussa havainnot keskistetyiksi. Mikäli tarve vaatiin, merkitään kuitenkin $\mathbf{y}_i - \bar{\mathbf{y}}$.

Taulukossa 3.1 on ASANAT-aineiston 20 oppilaan arvosat, arvosanojen aritmeettiset keskiarvot w_{ka}, ensimmäinen pääkomponentti z_1 ja siitä muodostettu painotettu keskiarvo w_1. Huomattakoon, että pääkomponentit on laskettu koko aineiston (30 oppilasta) kovarianssimatriisista. Taulukon kolme viimeistä muuttujaa ovat arvosanojen lineaarisia yhdisteitä, joiden kertoimille on asetettu tiettyjä reunaehtoja. Keskiarvomuuttujien w_{ka} ja w_1 kertoimien summa on yksi, kun taas pääkomponenttien kertoimien neliöiden summa on yksi (kerroinvektorin pituus = 1).

Taulukko 3.1. ASANAT-aineiston 20 oppilaan arvosat, arvosanojen aritmeettiset keskiarvot w_{ka}, ensimmäinen pääkomponentti z_1 ja painotettu keskiarvo w_1.

Engl.	Ruotsi	Alg.	Geom.	Fys.	Saksa	w_{ka}	z_1	w_1
7	9	8	8	7	7	7.83	0.14	7.73
8	8	8	10	9	8	8.67	2.27	8.66
8	9	9	6	8	7	8.00	0.18	7.75
7	8	8	6	6	8	7.33	-1.60	6.98
10	9	8	6	7	9	8.33	0.68	7.97
9	7	9	9	9	7	8.50	2.08	8.58
9	9	9	7	6	8	8.17	0.71	7.98
8	8	10	8	8	8	8.50	1.54	8.34
9	9	9	7	10	9	9.00	2.21	8.64
10	9	8	7	7	8	8.33	1.14	8.17
7	7	7	4	7	8	6.83	-3.19	6.29
5	6	8	6	7	6	6.50	-3.20	6.28
8	7	7	6	8	9	7.67	-1.04	7.22
7	7	8	6	6	7	7.00	-2.09	6.77
9	8	7	9	7	7	8.00	0.95	8.09
9	9	9	8	8	7	8.50	1.85	8.48
7	8	7	5	6	8	7.00	-2.59	6.55
9	10	10	8	8	8	9.00	2.73	8.86
9	8	6	6	9	7	7.67	-0.53	7.45
10	10	10	8	8	9	9.33	3.35	9.13

Tarkastellaan ensin kahden muuttujan lineaarista yhdistettä

$$\begin{aligned} z &= a_1 y_1 + a_2 y_2 \\ &= \mathbf{a}'\mathbf{y}. \end{aligned}$$

Muuttujan z varianssi on

$$\begin{aligned} s_z^2 &= a_1^2 s_1^2 + a_2^2 s_2^2 + 2a_1 a_2 s_{12} \\ &= (a_1, a_2) \begin{pmatrix} s_1^2 & s_{12} \\ s_{21} & s_2^2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \\ &= \mathbf{a}'\mathbf{S}\mathbf{a}, \end{aligned}$$

missä siis

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} s_1^2 & s_{12} \\ s_{21} & s_2^2 \end{pmatrix}, \quad s_{12} = s_{21},$$

s_{11} on y_1:n varianssi ja s_{22} on y_2:n varianssi. Pääkomponenttianalyysissa etsitään sellaisia kertoimien a_1 ja a_2 arvoja, että z:n varianssi s_z^2 maksimoituu.

Varianssia s_z^2 voidaan kasvattaa rajatta, mikäli a_1:n ja a_2:n arvoille ei panna ylärajaa. Siksi kerroinvektorin pituus normeerataan ykköseksi:

$$\|\mathbf{a}\| = \sqrt{a_1^2 + a_2^2} = 1.$$

Esimerkiksi

$$z = 0.6y_1 + 0.8y_2$$

on normeerattu lineaarinen yhdiste, sillä $0.6^2 + 0.8^2 = 1$. Tämän rajoitteen vallitessa maksimointitehtävä

$$\max_{\|\mathbf{a}\|=1} s_z^2 = \max_{\|\mathbf{a}\|=1} \mathbf{a}'\mathbf{S}\mathbf{a}$$

on ratkaistavissa yksikäsitteisesti.

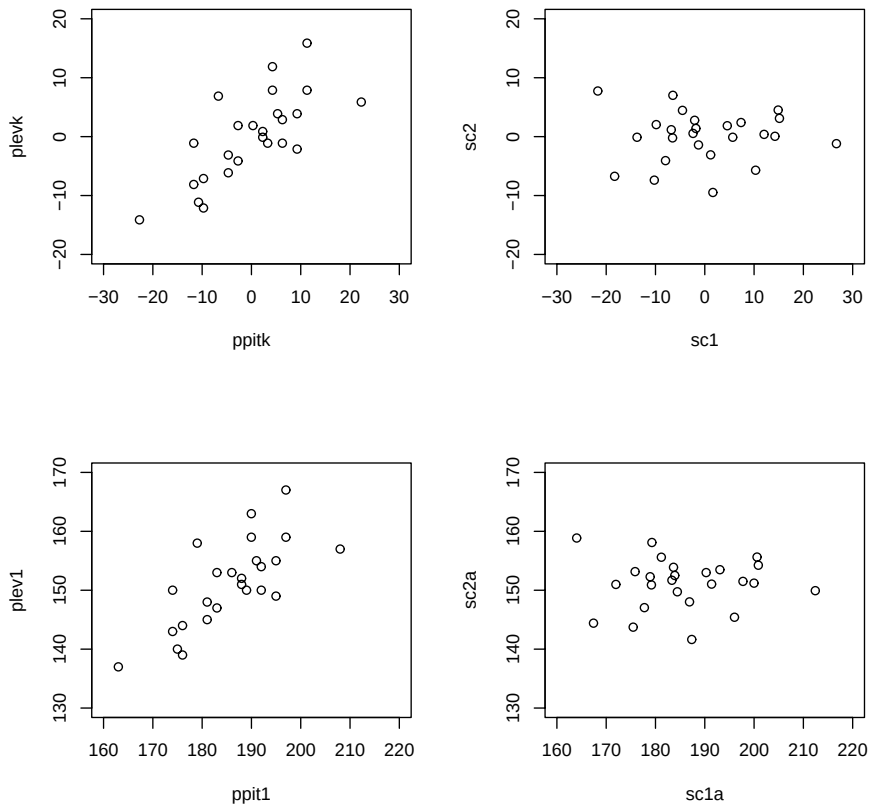
Esimerkki 3.1 Tarkastellaan havaintoaineistoa `Pojat1`, jossa on annettu 25 perheen 1. pojan pään pituus ja leveys (mm) täysikasvuisena. Määritetään tästä aineistosta pääkomponentit. Lasketaan ensin kovarianssimatriisin ominaisarvot ja ominaisvektotrit.

```
S<-cov(pojat1[,1:2])
S.eigen<-eigen(S,T)
P<-S.eigen$vectors
P
           plev1      ppit1
ppit1 -0.8249295  0.5652357
plev1 -0.5652357 -0.8249295

l<-S.eigen$values
l
[1] 131.5183  18.1350
```

Matriisin $\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2)$ ensimmäinen sarake $\mathbf{p}_1$ on suurimapaan ominaisarvoon $\lambda_1 = 131.5183$ liittyvä ominaisvektori ja $\mathbf{p}_2$ on ominaisarvoon $\lambda_2 = 18.1350$ liittyvä ominaisvektori. Matriisi $\mathbf{A}$ on ortogonaalinen, eli $\mathbf{p}_1'\mathbf{p}_1 = \mathbf{p}_2'\mathbf{p}_2 = 1$ ja $\mathbf{p}_1'\mathbf{p}_2 = 0$. Merkitään nyt $\mathbf{A} = \mathbf{P}', y_1$ =plev1, y_2 =ppit1 ja $\mathbf{y} = (y_1, y_2)$. Silloin pääkomponentit määritellään yhtälöllöllä $\mathbf{z} = \mathbf{A}\mathbf{y}$ tai

$$\begin{aligned} z_1 &= \mathbf{a}_1'\mathbf{y} \\ z_2 &= \mathbf{a}_2'\mathbf{y}, \end{aligned}$$

missä z_1 on 1. pääkomponentti ja z_2 on 2. pääkomponentti. Muuttujat z_1 ja z_2 sisältävät ns. pääkomponettipisteet. Esimerkiksi $z_{i1} = \mathbf{a}_1'\mathbf{y}_i$ on i. havainnon pistemäärä (arvo) 1. pääkomponentilla ja $z_{i2} = \mathbf{a}_2'\mathbf{y}_i$ on i. havainnon pistemäärä (arvo) 2. pääkomponentilla. Kertoimien $\mathbf{a}_i$ alkioita kutsutaan usein



Kuva 3.1: Pään pituuden ja leveyden sekä niistä laskettujen pääkomponenttien pisteparvet. Aineistona 25 perheen 1. poikalapset.

latauksiksi. Esimerkiksi a_{11} on muuttujan `ppit1` ja a_{12} muuttujan `plev1` lataus 1. pääkomponentilla.

Tehdään nyt pääkomponenttianalyysi komennolla `princom`.

□

Akselien kierto saadaan aikaan kertomalla havaintovektorit $\mathbf{y}_1$, $\mathbf{y}_2$, ..., $\mathbf{y}_n$ *ortogonaalisella matriisilla* $\mathbf{A}$:

$$\mathbf{z}_i = \mathbf{A}\mathbf{y}_i.$$

Koska $\mathbf{A}$ on ortogonaalinen ($\mathbf{A}'\mathbf{A} = \mathbf{I}$), pisteet $\mathbf{y}_i$ ja $\mathbf{z}_i$ ovat samalla etäisyydellä origosta:

$$\mathbf{z}_i'\mathbf{z}_i = (\mathbf{A}\mathbf{y}_i)'(\mathbf{A}\mathbf{y}_i) = \mathbf{y}_i'\mathbf{A}'\mathbf{A}\mathbf{y}_i = \mathbf{y}_i'\mathbf{y}_i.$$

Nyt halutaan löytää sellainen ortogonaalinen matriisi $\underset{p\times p}{\mathbf{A}}$, että muunnoksessa $\mathbf{z} = \mathbf{A}\mathbf{y}$ syntyvät muuttujat $z_1, z_2, \ldots, z_p$ ovat *korreloimattomia*. Silloin $\mathbf{z}$:n otoskovarianssimatriisi on muotoa

$$\mathbf{S}_z = \begin{pmatrix} s_{z_1}^2 & 0 & \ldots & 0 \\ 0 & s_{z_1}^2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & s_{z_p}^2 \end{pmatrix}.$$

Koska toisaalta $\mathbf{S}_z = \mathbf{A}\mathbf{S}\mathbf{A}'$, niin

$$\mathbf{A}\mathbf{S}\mathbf{A}' = \begin{pmatrix} s_{z_1}^2 & 0 & \ldots & 0 \\ 0 & s_{z_1}^2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & s_{z_p}^2 \end{pmatrix}, \tag{3.2.1}$$

missä $\mathbf{S}$ on $\mathbf{y}$:n otoskovarianssimatriisi. On siis löydettävä ortogonaalinen matriisi, joka diagonalisoi $\mathbf{S}$:n. Symmetristen matriisien ominaisarvotehtävän ratkaisua koskevien tulosten perusteella tiedetään, että $\mathbf{S}$:n diagonalisoivan ortogonaalisen matriisin $\mathbf{A}$ rivit ovat $\mathbf{S}$:n normeerattuja ominasivektoreita. Matriisin $\mathbf{A}\mathbf{S}\mathbf{A}'$ diagonaalielementit yhtälössä (3.2.1) ovat $\mathbf{S}$:n ominaisarvot. Pääkomponentit ovat siis alkuperäisten muuttujien $\mathbf{y}' = (y_1, y_2, \ldots, y_p)$ lineaariset yhdisteet

$$\begin{aligned} z_1 &= \mathbf{a}_1'\mathbf{y} \\ z_2 &= \mathbf{a}_2'\mathbf{y} \\ &\vdots \\ z_p &= \mathbf{a}_p'\mathbf{y}, \end{aligned} \tag{3.2.2}$$

joiden kertoimet $\mathbf{a}_1$, $\mathbf{a}_2$, ..., $\mathbf{a}_p$ ovat $\mathbf{S}$:n normeerattut ominaisvektorit ja ominaisarvot $\lambda_1, \lambda_2, \ldots, \lambda_p$ ovat pääkomponenttien z_1, z_2, ..., z_p varianssit:

$$s^2_{z_i} = \lambda_i.$$

Ominaisarvot ja ominaisvektorit saadaan ratkaisemalla yhtälö

$$(\mathbf{S} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0}.$$

Suurimpaan ominaisarvoon λ_1 liittyvä ominaisvektori $\mathbf{a}_1$ on siis 1. pääkomponentin $z_1 = \mathbf{a}_1'\mathbf{y}$ kerroinvektori.

3.3 Pääkomponenttipistemäärät

Pääkomponenttien z_i, $i = 1, 2, \ldots, p$ arvoja sanotaan myös pääkomponenttipistemääriksi. Pääkomponentit ovat keskistettyinä muuttujia ja ne ovat muotoa

$$\mathbf{z}_i = \mathbf{a}_i'(\mathbf{y} - \bar{\mathbf{y}}),$$

missä $\bar{\mathbf{y}} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_p)'$. Silloin esimerkiksi k. tilastoyksikön arvo 1. pääkomponentilla on

$$z_{k1} = \mathbf{a}_1'(\mathbf{y}_k - \bar{\mathbf{y}}) = \mathbf{a}_1'\mathbf{y}_k - \mathbf{a}_1'\bar{\mathbf{y}}.$$

Esimerkki 3.2 Tehdään nyt aineistosta `pojat1` pääkomponenttianalyysi komennolla `princom`.

```
library(help=Mva)
pojat.pka<-princomp(pojat1)
pojat.pka
Call:
princomp(x = pojat1)

Standard deviations:
   Comp.1    Comp.2
11.236441  4.172481

 2  variables and  25 observations.

summary(pojat.pka)  # Kokeile myös plot(pojat.pka)

Importance of components:
                          Comp.1    Comp.2
Standard deviation     11.23644 4.1724812
Proportion of Variance  0.87882 0.1211801
Cumulative Proportion   0.87882 1.0000000
```

Peruskomennot `princomp(pojat1)` ja `summary(princomp(pojat1))` tulostavat vain analyysin keskeisimmän asian: pääkomponenttien hajonnat ja pääkomponentin varianssin suhteellisen osuuden kokonaisvarianssista. Komponentit esitetään järjestyksessä niiden varianssien suuruuden mukaan. Analyysin tulosta on kuitenkin mahdollista tarkastella lämmin. Katsotaan komennon `names(pojat.pka)` avulla, mitä analyysiin liittyvän laskennan tuloksia on käytettävissä.

```
names(pojat.pka)
[1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"   "call"
pojat.pka$loadings

Loadings:
      Comp.1 Comp.2
ppit1 -0.825  0.565
plev1 -0.565 -0.825

pojat.pka$scores[1:5,]

       Comp.1      Comp.2
  1 -6.548742 -0.21628175
  2 -6.457046  6.99423778
  3  5.657202 -0.09413264
  4  1.181165 -3.08830848
  5 12.042792  0.37940663
```

Edellä on tulostettu esimerkin vuoksi lataukset eli pääkomponenttien kerroinvektorit $\mathbf{a}_1 = (-0.825, -0.565)'$ ja $\mathbf{a}_2 = (0.565, -0.825)'$ sekä pääkomponenttipistemäärät (5 ensimmäistä havaintoa). Pääkomponenttipistemäärien korrelaatiodiagrammit saadaan komennolla `plot(pojat.pka$scores)`. □

Esimerkki 3.3 Tarkastellaan nyt ARVOSANAT-aineistoa, jossa on 30 oppilaan kuuden oppiaineen arvosanat. Kovarianssimatriisi on

$$\mathbf{S} = \begin{pmatrix} 1.334 & 0.628 & 0.345 & 0.528 & 0.369 & 0.417 \\ 0.628 & 0.892 & 0.414 & 0.338 & 0.191 & 0.274 \\ 0.345 & 0.414 & 1.172 & 0.483 & 0.414 & 0.138 \\ 0.528 & 0.338 & 0.483 & 1.955 & 0.321 & -0.072 \\ 0.369 & 0.191 & 0.414 & 0.321 & 1.357 & -0.043 \\ 0.417 & 0.274 & 0.138 & -0.072 & -0.043 & 0.861 \end{pmatrix}.$$

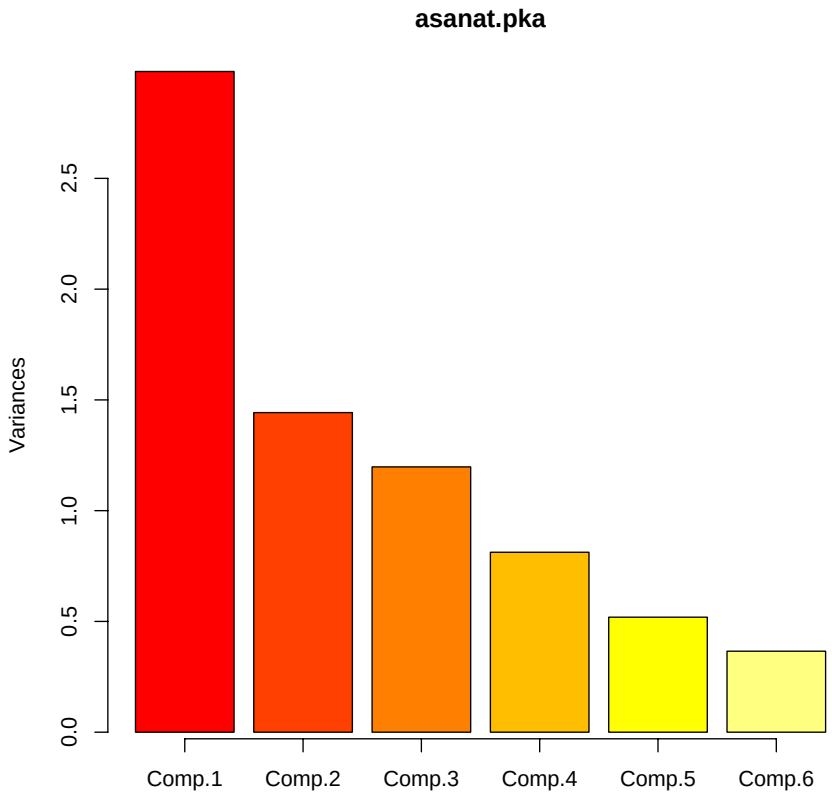
Pääkomponenttien varianssit ovat

	z_1	z_2	z_3	z_4	z_5	z_6
Varianssi	3.085	1.493	1.239	0.840	0.537	0.378
Prosentteina	40.746	19.710	16.360	11.097	7.092	4.995
Kumulatiivinen	40.746	60.456	76.816	87.913	95.005	100.000

Kokonaisvarianssi on

$$\sum_{i=1}^{6} s_i^2 = \sum_{i=1}^{6} \lambda_i = 7.572.$$

□



Kuva 3.2: `Asanat`-aineiston pääkomponenttien varianssit pylväsdiagrammina

3.4 Pääkomponenttien määrittäminen

Muuttujien $y_1, y_2, \ldots, y_p$ lineaarinen yhdiste on muotoa

$$z = \mathbf{a}'\mathbf{y} = a_1 y_1 + a_2 y_2 + \ldots + a_p y_p$$

ja

$$\bar{z} = a_1 \bar{y}_1 + \ldots + a_p \bar{y}_p$$

sekä z:n otosvarianssi

$$s_z^2 = \mathbf{a}'\mathbf{S}\mathbf{a},$$

missä $\bar{y}_1, \ldots, \bar{y}_p$ ovat alkuperäisten havaintojen otoskeskiarvoja ja $\underset{p\times p}{\mathbf{S}}$ otoskorrelaatiomatriisi. Olkoon $v = \mathbf{b}'\mathbf{y}$ jokin toinen muuttujien $y_1, y_2, \ldots, y_p$ lineaarinen yhdiste. Silloin muuttujien z ja v otoskovarianssi on

$$s_{zv} = \mathbf{a}'\mathbf{S}\mathbf{b}.$$

3.4.1 Otoksesta laskettu 1. pääkomponentti

Ensimmäinen pääkomponentti

$$\begin{aligned} z_1 &= a_{11} y_1 + a_{12} y_2 + \ldots + a_{1p} y_p \\ &= \mathbf{a}_1' \mathbf{y} \end{aligned}$$

on se muuttujien $y_1, \ldots, y_p$ lineaarinen yhdiste, jonka otosvarianssi on suurin. Kerroinvektorin pituus on kiinnitetty ykköseksi ($\mathbf{a}_1'\mathbf{a}_1 = 1$). Ensimmäinen pääkomponentti toteuttaa siis epäyhtälön

$$\mathrm{var}(z_1) = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1 \geq \mathrm{var}(z) = \mathbf{a}'\mathbf{S}\mathbf{a},$$

olipa lineaarisen yhdisteen $z = \mathbf{a}'\mathbf{y}$ ykkösen pituinen kerroinvektori $\mathbf{a}$ valittu miten tahansa.

3.4.2 1. pääkomponentti

Toinen ottoksesta laskettu pääkomponentti on se muuttujien $y_1, \ldots, y_p$ lineaarinen yhdiste

$$\begin{aligned} z_2 &= a_{21} y_1 + a_{22} y_2 + \ldots + a_{2p} y_p \\ &= \mathbf{a}_2' \mathbf{y}, \end{aligned}$$

jonka otosvarianssi on suurin ensimmäisen pääkomponentin kanssa korreloimattomien lineaaristen yhdisteiden joukossa. Toinen pääkomponentti toteuttaa siis epäyhtälön

$$\mathrm{var}(z_2) = \mathbf{a}_2'\mathbf{S}\mathbf{a}_2 \geq \mathrm{var}(z) = \mathbf{a}'\mathbf{S}\mathbf{a}$$

ehdolla, että $\text{cov}(z_1, z_2) = \mathbf{a}_1'\mathbf{S}\mathbf{a}_2 = 0$ ja $\mathbf{a}'\mathbf{a} = \mathbf{a}_1'\mathbf{a}_1 = \mathbf{a}_2'\mathbf{a}_2 = 1$. Toinen pääkomponentti selittää mahdollisimman paljon siitä vaihtelusta, joka jää ensimmäiseltä pääkomponentilta selittämättä. Vastaavasti i. pääkomponentti

$$z_i = a_{i1}y_1 + a_{i2}y_2 + \ldots + a_{ip}y_p = \mathbf{a}_i'\mathbf{y}$$

selittää mahdollisimman paljon siitä kokonaisvarianssista, joka jää $i-1$:ltä ensimmäiseltä pääkomponentilta selittämättä.

3.5 Pääkomponenttien laskeminen

Voidaan osoittaa, että

$$\max_{\mathbf{a}'\mathbf{a}=1} \mathbf{a}'\mathbf{S}\mathbf{a} = \lambda_1 > 0$$

on $\mathbf{S}$:n suurin ominaisarvo ja maksimi saavutetaan, kun $\mathbf{a} = \mathbf{a}_1$ on suurimpaan ominaisarvoon liittyvä normeerattu ominaisvektori. Näin siis 1. pääkomponentti on

$$z_1 = \mathbf{a}_1'\mathbf{y}.$$

Pääkomponentit voidaan laskea ratkaisemalla ominaisarvotehtävä

$$\mathbf{S}\mathbf{a} = \lambda\mathbf{a}.$$

Merkitään $\mathbf{S}$:n ominaisarvoja $\lambda_1, \lambda_2, \ldots, \lambda_p$ (suuruusjärjestyksessä) ja vastaavia ominaisvektoreita $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$. Silloin i. pääkomponentti on

$$z_i = \mathbf{a}_i\mathbf{y} = a_{i1}y_1 + a_{i2}y_2 + \ldots + a_{ip}y_p$$

ja sen varianssi

$$s_{z_i}^2 = \mathbf{a}_i'\mathbf{S}\mathbf{a}_i = \lambda_i.$$

on $\mathbf{S}$:n i. ominaisarvo.

Ominasiarvoyhtälön

$$\mathbf{S}\mathbf{a} = \lambda\mathbf{a} \Leftrightarrow |\mathbf{S} - \lambda\mathbf{I}_p| = 0$$

ratkaisuna saadaan p ei-negatiivista ominaisarvoa λ_1, λ_2, $\ldots$, λ_p ja niitä vastaavat p ortonormaalista ominaisvektoria $\mathbf{a}_1$, $\mathbf{a}_2$, $\ldots$, $\mathbf{a}_p$. Ominaisvektorien ortonormaalisuus tarkoittaa, että $\mathbf{a}_i^{'}\mathbf{a}_j = 0$, kun $i \neq j$ ja $\mathbf{a}_i^{'}\mathbf{a}_i = 1$. Siis esimerkiksi

$$\mathbf{a}_1^{'}\mathbf{S}\mathbf{a}_1 = \lambda_1 = s_{z_1}^2$$

on 1. pääkomponentin varianssi, ja yleisesti

$$\mathbf{a}_i^{'}\mathbf{S}\mathbf{a}_i = \lambda_1 = s_{z_i}^2.$$

Kuten edellä on todettu, i. pääkomponentin määrittävän lineaarisen yhdisteen kertoimet ovat kovarianssimatriisin $\mathbf{S}$ i. ominaisvektorin $\mathbf{a}_i$ alkioita.

Merkitään

$$\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p)'.$$

Silloin saadaan

$$\mathbf{ASA}' = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_p \end{pmatrix} = \mathbf{\Lambda} \Leftrightarrow \quad \mathbf{S} = \mathbf{A}'\mathbf{\Lambda}\mathbf{A}.$$

Matriisi $\mathbf{\Lambda}$ on siis diagonaalimatriisi, jonka diagonaalialkiot ovat $\mathbf{S}$:n ominaisarvoja. Kokonaisvarianssi on

$$\operatorname{tr} \mathbf{S} = s_{11} + s_{22} + \ldots + s_{pp} = \operatorname{tr}(\mathbf{A}'\mathbf{\Lambda}\mathbf{A}) = \operatorname{tr} \mathbf{\Lambda} = \lambda_1 + \lambda_2 + \ldots + \lambda_p.$$

Suhdelukua

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}, \quad k = 1, 2, \ldots, p$$

käytetään mittaamaan k. pääkomponentin selittämää osuutta kokonaisvaihtelusta. Pyrkimyksenä on, että suurehkoilla p:n arvoilla pari kolme ensimmäistä komponenttia selittäisi 80–90 % kokonaisvaihtelusta.

3.6 Tulkinta korrelaatioiden avulla

Pääkomponenttien $z_1, z_2, \ldots, z_p$ ja alkuperäisten muuttujien $y_1, y_2, \ldots, y_p$ väliset korrelaatiot saattavat olla tulkinnan kannalta hyödyllisiä. Määritelmän mukaan

$$\begin{aligned} r_{z_i, y_k} &= \frac{\operatorname{cov}(z_i, y_k)}{\sqrt{\operatorname{var}(z_i)} \cdot \sqrt{\operatorname{var}(y_k)}} \\ &= \frac{\operatorname{cov}(\mathbf{a}_i'\mathbf{y}, \mathbf{e}_k'\mathbf{y})}{\sqrt{\lambda_i} \cdot \sqrt{s_{kk}}} \\ &= \frac{\mathbf{a}_i'\mathbf{S}\mathbf{e}_k}{\sqrt{\lambda_i} \cdot \sqrt{s_{kk}}} \\ &= \frac{\lambda_i \mathbf{a}_i'\mathbf{e}_k}{\sqrt{\lambda_i}\ s_k} \\ &= \frac{\sqrt{\lambda_i}\ a_{ki}}{s_k}, \end{aligned}$$

missä $\mathbf{e}_k = (0, \ldots, 0, 1, 0, \ldots, 0)'$ on k. kantavektori ja $\operatorname{cov}(z_i, y_k)'$ merkitsee z_i:n ja y_k:n välistä kovarianssia.

Esimerkki 3.4 Lasketaan `pojat1`-aineistosta alkuperäisten muuttujien ja pääkomponenttien väliset korrelaatiot.

```
cor(pojat.pka$scores,pojat1)
            ppit1      plev1
Comp.1 -0.9691225 -0.8791906
Comp.2  0.2465796 -0.4764703
```

Huomataan, että 1. komponetti korreloi voimakkaimmin muuttujan `ppit1` kanssa. Tosin myös muuttuja `plev1` on tärkeä muuttuja 1. komponetissa. 2. komponetin tärkein muuttuja on `plev1`.

Ohessa muuttujien ja pääkomponenttien väliset korrelaatiot `asanat`-aineistosta.

```
cor(asanat.pka$scores,asanat)

             engl     ruotsi          alg        geom         fys       saksa
Comp.1 -0.7385948 -0.6530032 -0.64285889 -0.74106056 -0.51799531 -0.25194077
Comp.2  0.4469060  0.4143482  0.05458564 -0.59205963 -0.01413796  0.66763357
Comp.3  0.1838501  0.1569885 -0.25469752  0.28767344 -0.79678061  0.28941042
Comp.4 -0.3282820  0.1284623  0.69734869 -0.06427443 -0.27415837  0.04177957
Comp.5 -0.1431818 -0.4004759  0.01730113  0.10455859  0.10770589  0.61831494
Comp.6 -0.3044367  0.4476310 -0.17968722  0.04974864  0.09928880  0.15157171

asanat.pka$loadings

Loadings:
       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
engl   -0.486  0.423  0.191 -0.414 -0.226 -0.572
ruotsi -0.351  0.320  0.133  0.132 -0.516  0.687
alg    -0.396        -0.248  0.824        -0.316
geom   -0.590 -0.678  0.361         0.200  0.113
fys    -0.344        -0.834 -0.348  0.171  0.188
saksa  -0.133  0.507  0.241         0.783  0.229
```

Nähdään, että 1. pääkomponentissa kaikki muuttujat, saksaa lukuunottamatta, ovat jokseenkin yhtä tärkeitä. Toisessa komponentissa ovat taas tärkeimpiä englanti, ruotsi, geometria ja saksa. Lataukset heijastelevat samaa kaavaa: 1. komponentin lataukset painottavat viittä ensimmäistä muuttujaa melko tasaisesti, vain saksalla on selvästi pienempi lataus. □

Esimerkki 3.5 Tarkastellaan nyt aineistoa `Osakkeet`. Siinä tarkastellaan viiden eri osakkeen tuottoa New Yorkin pörssissä tammikuun 1975 ja joulukuun 1976 välisenä aikana. Tuotto on määritellään siten, että perjantain hinnasta pörssin sulkeutuessa vähennetään vastaava edellisen perjantain hinta ja erotus jaetaan edellisen perjantain hinnalla. Aineistossa on 100:n viikon tuotot.

```
osakkeet[1:5,]
  Allied.Ch    DuPont U.Carbide     Exxon    Texaco
1  0.000000  0.000000  0.000000  0.039473  0.000000
2  0.027027 -0.044855 -0.003030 -0.014466  0.043478
3  0.122807  0.060773  0.088146  0.086238  0.078124
4  0.057031  0.029948  0.066808  0.013513  0.019512
5  0.063670 -0.003793 -0.039788 -0.018644 -0.024154

diag(cov(osakkeet))
   Allied.Ch       DuPont    U.Carbide        Exxon       Texaco
0.0016299269 0.0012293651 0.0015560763 0.0008023323 0.0007587370
```

Edellä on tulostettu esimerkiksi viiden ensimmäisen viikon tuotot ja laskettu muuttujien varianssit. Nähdään, että varianssit ovat suurinpiirtein samaa kertaluokkaa ja kaikki muuttujat on selkeästi mitattu samalla yksiköllä. Hajontaa onkin tässä perusteltua käyttää mittayksikkönä.

```
cor(osakkeet)
           Allied.Ch    DuPont U.Carbide     Exxon    Texaco
Allied.Ch 1.0000000 0.5769308 0.5086555 0.3867206 0.4621781
DuPont    0.5769308 1.0000000 0.5983817 0.3895188 0.3219545
U.Carbide 0.5086555 0.5983817 1.0000000 0.4361014 0.4256266
Exxon     0.3867206 0.3895188 0.4361014 1.0000000 0.5235293
Texaco    0.4621781 0.3219545 0.4256266 0.5235293 1.0000000
```

Osakkeiden tuotot korreloivat varsin selvästi keskenään. Sen sijaan peräkkäisten 100:n viikon tuotot näyttävät oleva jokseenkin riippumattomat toisistaan. Seuraavassa on laskettu Allied Chemicalsin osakkeen viikkotuottojen autorrelaatiot aina viiden viikon viiveeseen asti. Autokorrelaatiot ovat varsin pieniä.

```
library(ts)
osakk.akor<-acf(Allied.Ch,lag.max = 5,plot=F)
> osakk.akor$acf
, , 1

             [,1]
[1,]  1.00000000
[2,] -0.01670283
[3,] -0.04418965
[4,]  0.09274225
[5,] -0.04325547
[6,] -0.01700809
```

□

3.7 Pääkomponenttien laskeminen korrelaatiomatriisista

Esimerkin 3.5 aineistossa havaintoarvo y_{ij} on j. yhtiön osakkeen tuotto viikolla i New Yorkin pörssissä. Esimerkissä muuttujien lukumäärä p on 5 ja yhtiöt ovat: Allied Chemical (1), DuPont (2), Union Carbide (3), Exxon (4) ja Texaco (5). Jokaisesta yhtiöstä on käytettävissä $n = 100$ havaintoa, joten havaintomatriisin koko on 100×5.

Olkoot v_1, v_2, ..., v_p standardoituja muuttujia ja

$$v_{kj} = \frac{y_{kj} - \bar{y}_j}{s_j}$$

missä $j = 1, 2, \ldots, p$ ja $k = 1, 2, \ldots, n$ standardoituja havaintoja. Jos $\mathbf{v} = (v_1, v_2, \ldots, v_p)'$ on standardoitujen muuttujien vektori ja $\mathbf{V}$ standardoitujen havaintojen havaintomatriisi, niin korrelaatiomatriisi voidaan kirjoittaa muodossa

$$\underset{p\times p}{\mathbf{R}} = \frac{1}{n-1}\mathbf{V}'\mathbf{V}.$$

Merkitään ominaisarvoja ja ominaisvektoreita samoin kuin kovarianssimatriisin tapauksessa. $\mathbf{R}$:n ominaisarvot ovat λ_1, λ_2, ..., λ_p ja ominaisvektorit $\mathbf{a}_1$, $\mathbf{a}_2$, ..., $\mathbf{a}_p$. Huomattakoon, että otoskovarianssimatriisista $\mathbf{S}$ ja otoskorrelaatiomatriisista $\mathbf{R}$ lasketut pääkomponentit eivät ole samoja.

Otoskorrelaatiomatriisista laskettu i. pääkomponentti on

$$z_i = \mathbf{a}_i'\mathbf{v} = a_{i1}v_1 + a_{i2}v_2 + \ldots + a_{ip}v_p, \quad i = 1, 2, \ldots, p,$$

missä $\mathbf{a}_i$ on $\mathbf{R}$:n i. ominaisarvoon liittyvä normeerattu ominaisvektori. Samalla tavalla kuin otoskovarianssimatriisin tapauksessa

$$\mathrm{var}(z_i) = s_{z_i}^2 = \mathbf{a}_i'\mathbf{R}\mathbf{a}_i = \lambda_i, \quad i = 1, \ldots, p$$

ja

$$\mathrm{cov}(z_i, z_j) = 0, \quad i \neq j.$$

Standardoitujen muuttujien kokonaisvarianssi on

$$\mathrm{tr}(\mathbf{R}) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = p$$

ja korrelaatiokerroin

$$r_{z_i, v_k} = a_{ki}\sqrt{\lambda_i}, \quad i = 1, 2, \ldots, p.$$

Esimerkki 3.6 Tehdään nyt pääkomponenttianalyysi aineiston `Osakkeet` korrelaatiomatriisin perusteella.

```
osakkeet.pka<-princomp(osakkeet,cor=T)
summary(osakkeet.pka)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
Standard deviation     1.6901150 0.8995104 0.7348756 0.67182459 0.58566358
Proportion of Variance 0.5712978 0.1618238 0.1080084 0.09026966 0.06860037
Cumulative Proportion  0.5712978 0.7331216 0.8411300 0.93139963 1.00000000
```

Ensimmäinen pääkomponentti selittää noin 57% kokonaisvaihtelusta.

```
osakkeet.pka$loadings

Loadings:
          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
Allied.Ch -0.464  0.241 -0.613 -0.381  0.453
DuPont    -0.457  0.509  0.178 -0.211 -0.675
U.Carbide -0.470  0.261  0.337  0.664  0.396
Exxon     -0.422 -0.525  0.539 -0.473  0.179
Texaco    -0.421 -0.582 -0.434  0.381 -0.387

cor(osakkeet,osakkeet.pka$scores)
              Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
Allied.Ch -0.7834383  0.2166542 -0.4507341 -0.2562064  0.2654851
DuPont    -0.7725126  0.4579390  0.1307274 -0.1419682 -0.3953120
U.Carbide -0.7943196  0.2343861  0.2476899  0.4461625  0.2317504
Exxon     -0.7126819 -0.4724837  0.3961083 -0.3176426  0.1050956
Texaco    -0.7120944 -0.5237309 -0.3186520  0.2561130 -0.2269242
```

Esimmäisessä pääkomponentissa kaikkia muuttujia painotetaan yhtä paljon ja kaikki muuttujat ovat siinä jokseenkin yhtä tärkeitä. Se on siis eräänlainen markkinakomponentti. Kun pörssi on nousussa, nämä kaikki osakkeet tuottavat. □

3.8 Populaation pääkomponentit

Olkoon $\mathbf{y} = (y_1, y_2, \ldots, y_p)'$ satunnaisvektori, jonka kovarianssimatriisi

$$\mathrm{cov}(\mathbf{y}) = \mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{pmatrix},$$

missä siis $\mathrm{cov}(y_i, y_j) = \sigma_{ij}$. Jos $\mathbf{\Sigma}$ tunnetaan, niin populaation pääkomponentit voidaan laskea aivan samalla tavalla kuin otoskovarianssimatriisista

S. Korvataan siis edellä esitetyissä kaavoissa **S** teoreettisella kovarianssimatriisilla $\boldsymbol{\Sigma}$. Merkitään $\boldsymbol{\Sigma}$:n ominaisarvoja ja ominaisvektoreita

$$\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_p \quad \text{ja} \quad \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_p.$$

Populaation pääkomponentit ovat

$$\begin{aligned} z_1 &= \boldsymbol{\alpha}_1' \mathbf{y} = \alpha_{11} y_1 + \alpha_{12} y_2 + \ldots + \alpha_{1p} y_p \\ z_2 &= \boldsymbol{\alpha}_2' \mathbf{y} = \alpha_{21} y_1 + \alpha_{22} y_2 + \ldots + \alpha_{2p} y_p \\ &\vdots \\ z_p &= \boldsymbol{\alpha}_p' \mathbf{y} = \alpha_{p1} y_1 + \alpha_{p2} y_2 + \ldots + \alpha_{pp} y_p \end{aligned}$$

ja

$$\begin{aligned} \operatorname{var} z_i &= \gamma_i \\ \operatorname{cov}(z_i, z_j) &= 0, \quad i \neq j \end{aligned} \tag{3.8.3}$$

sekä

$$\rho_{z_i, y_k} = \frac{\alpha_{ki}\sqrt{\gamma_i}}{\sqrt{\sigma_{kk}}} = \frac{\alpha_{ki}\sqrt{\gamma_i}}{\sigma_k}.$$

Käytännössä $\boldsymbol{\Sigma}$ on tuntematon ja sen estimaattina käytetään otoskovarianssimatriisia **S**.

3.9 Mittayksikön vaikutus pääkomponentteihin

Kovarianssimatriisiin perustuvan pääkomponenttianalyysin tulos poikkeaa usein perusteellisesti korrelaatiomatriisiin perustuvan analyysin tuloksesta. On aina tehtäväkohtaisesti päätettävä, kumpi lähestymistapa valitaan. On siis muistettava, että kyseessä on kaksi täysin eri analyysia. Muuttujien mittayksikön valinta vaikuttaa ratkaisevasti pääkomponenttianalyysin tulokseen. Muuttujat on yleensä järkevää standardoida, jos niiden varianssit poikkeavat radikaalisti toisistaan tai muuttujat eivät ole yhteismitallisia. Muuttuja, jonka varianssi on paljon suurempi muiden muuttujien varianssia, saattaa dominoida tuloksia pelkästään tämän ominaisuuden perusteella.

Esimerkki 3.7 Määritetään pääkomponentit kovarianssimatriisista

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix}$$

ja siitä lasketusta korrelaatiomatriisista

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}.$$

Ominaisarvot ja ominaisvektorit kovarianssimatriisista $\boldsymbol{\Sigma}$ ovat

$$\begin{aligned} \gamma_1 &= 100.16 & \boldsymbol{\alpha}_1 &= (0.040, 0.999)' \\ \gamma_2 &= 0.84 & \boldsymbol{\alpha}_2 &= (0.999, -0.040)' \end{aligned}$$

Vastaavasti korrelaatiomatriisista $\boldsymbol{\rho}$ lasketut ominaisarvot ja ominaisvektorit ovat

$$\begin{aligned} \gamma_1 &= 1.4 & \boldsymbol{\alpha}_1 &= (0.707, 0.707)' \\ \gamma_2 &= 0.6 & \boldsymbol{\alpha}_2' &= (0.707, -0.707)' \end{aligned}$$

□

Esimerkki 3.8 Generoidaan 100 havaintoa normaalijakausta $\mathrm{N}(\mathbf{0}, \mathbf{I}_3)$, missä $\mathbf{I}_3$ on 3×3-identiteettimatriisi. Lasketaan havaintojen kovarianssimatriisista ominaisarvot.

```
X<-matrix(rnorm(300),nrow=100)
cov(X)
             [,1]         [,2]         [,3]
[1,]  0.86892072 -0.01302330 -0.04564921
[2,] -0.01302330  0.97154678  0.09172346
[3,] -0.04564921  0.09172346  1.02315548
eigen(cov(X))
$values
[1] 1.1011227 0.9080082 0.8544921
```

Populaation kovarianssimatriisin ominaisarvot ovat ykkösiä, mutta aineistosta lasketut estimaatit poikkeavat kohtalaisen paljon teoreettisista arvoista, vaikka otoskoko on 100. □

3.10 Montako pääkomponenttia tarvitaan?

Käytännössä pääkomponentit lasketaan otoksen kovarianssimatriisista $\mathbf{S}$ tai korrelaatiomatriisista. Otosvaihtelun vuoksi λ_i ja $\mathbf{a}_i$ poikkeavat vastaavista populaation suureista. Ominaisarvojen ja ominaisvektoreiden otosjakaumat ovat vaikeita johtaa.

Voidaan osoittaa, että suurilla n:n arvoilla likimain

$$\sqrt{n}(\lambda_i - \gamma_i) \sim N(0, 2\gamma_i^2), \quad i = 1, 2, \ldots, p$$

ja

$$\mathrm{cov}(\lambda_i, \lambda_j) \approx 0.$$

Suurten otosten $100(1-\alpha)\%$:n luottamusväli on

$$\frac{\lambda_i}{1 + z_{\frac{\alpha}{2}}\sqrt{2/n}} \leq \gamma_i \leq \frac{\lambda_i}{1 - z_{\frac{\alpha}{2}}\sqrt{2/n}},$$

missä $z_{\frac{\alpha}{2}}$ on normaalijakauman $(1-\frac{\alpha}{2})100\%$-piste. Samanaikaiset Bonferroni-luottamusvälit m:lle λ_i:lle saadaan korvaamalla $z_{\frac{\alpha}{2}}$-prosenttipisteet $z_{\frac{\alpha}{2m}}$:llä.

Ei ole olemassa mitään yleisesti hyväksyttyä menettelyä määrittää sopiva pääkomponenttien lukumäärä. Esimerkiksi seuraavia kriteereitä voidaan käyttää.

1) Valitaan pääkomponentit, joita vastaavat ominaisarvot ovat tilastollisesti merkitseviä.

2) Valitaan komponentit, joiden varianssit ovat suurempia kuin $\bar{\lambda} = \sum_{i=1}^{p} \lambda/p$.

3) Otetaan mukaan niin monta komponenttia, että merkittävä osa kokonaisvaihtelusta (esim. 80 %) saadaan selitettyä.

4) Etsitään "pienten" ja "suurten" ominaisarvojen välinen raja. Pieniä vastaavat komponentit jätetään pois.

Harjoituksia

1. Laske FIRMAT aineistosta pääkomponentit. Tulkitse tulos.

2. FIRMAT-aineisto. Lausu 1. pääkomponentin varianssi alkuperäisten muuttujien varianssien ja kovarianssien avulla. Mitkä termit selittävät suurimman osuuden varianssista?

3. Määritä pääkomponentit OSAKKEET-aineistosta (kurssimuutokset 100 viikon ajalta). Tulkitse tuloksia.

4. Määritä pääkomponenttien lausekkeet sekä kovarianssimatriisista

$$\mathbf{S} = \begin{pmatrix} 1 & 4 \\ 4 & 100 \end{pmatrix}$$

että korrelaatiomatriisista

$$\mathbf{R} = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}.$$

5. Laske pääkomponentit korrelaatiomatriisista

$$\mathbf{R} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}.$$

6. Kerro FIRMAT aineistossa tulot-muuttuja kymmenellä ja tee sitten pääkomponenttianalyysi kovarianssimatriisin perusteella. Vertaa tulosta alkuperäisten muuttujien pääkomponenttianalyysiin.

7. Tarkastele FIRMAT-aineistoa. Laske
(a) kovarianssimatriisi sekä
(b) kokonaisvarianssi ja yleistetty varianssi.
(c) Kirjoita pääkomponenttien lausekkeet sekä

(d) pääkomponenttien varianssit ja kokonaisvarianssi.

8. FIRMAT-aineisto. Piirrä
(a) muuttujien väliset korrelaatiodiagrammit sekä
(b) pääkomponenttien väliset korrelaatiodiagrammit.

9. Tee CENSUS-aineistolle korrelaatiomatriisiin perustuva pääkomponenttianalyysi. Paljonko 2 ensimmäistä pääkomponenttia selittävät kokonaisvaihtelusta? Vertaile tuloksia kovariansimatriisiin perustuvan analyysin tuloksiin. Mikä on mittayksikön vaikutus pääkomponentteihin?

10. Laske kovarianssimatriisin

$$\mathbf{\Sigma} = \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}$$

pääkomponentit. Paljonko 1. pääkomponentti selittää kokonaisvaihtelusta?

11. Tee POJAT-aineiston 1. poikien pään mitoille pääkomponenttianalyysi. Miten tulkitset tuloksen?

12. Tee POJAT-aineistosta myös 2. pojille pääkomponenttianalyysi. Vertaile 1. ja 2. poikien tuloksia. Tee analyysi myös koko aineistolle. Miten nyt tulos tulkitaan?

13. Tarkastele ASANAT-aineistossa erilaisia "hyvyysindeksejä" (aritmeettinen keskiarvo, pääkomponentit). Piirrä kahden ensimmäisen pääkomponentin korrelaatiodiagramma. Tulkitse tulos.

Taulukko 3.1: ASANAT-aineiston 20 oppilaan arvosat, arvosanojen aritmeettiset keskiarvot w_{ka}, ensimmäinen pääkomponentti z_1 ja painotettu keskiarvo w_1.

Engl.	Ruotsi	Alg.	Geom.	Fys.	Saksa	w_{ka}	z_1	w_1
7	9	8	8	7	7	7.83	0.14	7.73
8	8	8	10	9	8	8.67	2.27	8.66
8	9	9	6	8	7	8.00	0.18	7.75
7	8	8	6	6	8	7.33	-1.60	6.98
10	9	8	6	7	9	8.33	0.68	7.97
9	7	9	9	9	7	8.50	2.08	8.58
9	9	9	7	6	8	8.17	0.71	7.98
8	8	10	8	8	8	8.50	1.54	8.34
9	9	9	7	10	9	9.00	2.21	8.64
10	9	8	7	7	8	8.33	1.14	8.17
7	7	7	4	7	8	6.83	-3.19	6.29
5	6	8	6	7	6	6.50	-3.20	6.28
8	7	7	6	8	9	7.67	-1.04	7.22
7	7	8	6	6	7	7.00	-2.09	6.77
9	8	7	9	7	7	8.00	0.95	8.09
9	9	9	8	8	7	8.50	1.85	8.48
7	8	7	5	6	8	7.00	-2.59	6.55
9	10	10	8	8	8	9.00	2.73	8.86
9	8	6	6	9	7	7.67	-0.53	7.45
10	10	10	8	8	9	9.33	3.35	9.13

Aineistot

ASANAT
CENSUS
OSAKKEET

Luku 4

Multinormaalijakauma

4.1 Multinormaalijakauman tiheysfunktio

4.1.1 Normaalijakauman tiheysfunktio

Jos $y \sim N(\mu, \sigma^2)$, niin sen tiheysfunktio on

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}, \quad -\infty < y < \infty.$$

4.1.2 Multinormaalijakauma

Jos $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, niin $\mathbf{y}$:n tiheysfunktio on

$$g(\mathbf{y}) = \frac{1}{(\sqrt{2\pi})^p \sqrt{|\boldsymbol{\Sigma}|}} e^{-\Delta^2/2},$$

missä p on muuttujien lukumäärä ja termi

$$\Delta^2 = (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})$$

on $\mathbf{y}$:n ja $\boldsymbol{\mu}$:n välinen **Mahalanobiksen etäisyys**.

4.1.3 Yleistetty varianssi

Kovarianssimatriisin $\boldsymbol{\Sigma}$ determinantti $|\boldsymbol{\Sigma}|$ on muuttujien $y_1, y_2, \ldots, y_p$ yleistetty varianssi. Esimerkiksi

$$\begin{vmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{vmatrix} = \begin{vmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{vmatrix} = \sigma_1^2\sigma_2^2(1-\rho^2)$$

on 2–ulotteisen normaalijakauman yleistetty varianssi, kun $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$, $\mathrm{var}(y_i) = \sigma_i^2$, $i = 1, 2$ ja $\mathrm{cov}(y_1, y_2) = \sigma_{12}$.

4.1.4 Standardoitu normaalijakauma

Standardoidun normaalijakauman $z \sim N(0,1)$ tiheysfunktio on

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty$$

ja standardoidun p–ulotteisen normaalijakauman $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I}_p)$ tiheysfunktio on

$$g(\mathbf{z}) = \frac{1}{(\sqrt{2\pi})^p} e^{-\mathbf{z}'\mathbf{z}/2}$$

4.2 Multinormaalijakaumaa noudattavien satunnaismuuttujien ominaisuuksia

Olkoon $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

4.2.1 Lineaarinen yhdiste

(a) Jos $\mathbf{a} = (a_1, a_2, \ldots, a_p)$ on vakiovektori, niin

$$\mathbf{a}'\mathbf{y} = a_1 y_1 + a_2 y_2 + \ldots + a_p y_p \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}),$$

(b) Jos $\underset{q\times p}{\mathbf{A}}$ on täysasteinen vakiomatriisi, niin

$$\mathbf{A}\mathbf{y} \sim N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}').$$

4.2.2 Standardoidut muuttujat

Olkoon $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I}_p)$ ja $\underset{p\times p}{\boldsymbol{\Sigma}} = \mathbf{T}\mathbf{T}'$ positiivisesti definiitti matriisi. Silloin

$$\mathbf{y} = \mathbf{T}\mathbf{z} + \boldsymbol{\mu} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Jos vaaditaan, että $\mathbf{T}$ on positiivisesti definiitti yläkolmiomatriisi, niin $\boldsymbol{\Sigma} = \mathbf{T}\mathbf{T}'$ on ns. Choleskyn hajotelma. Vastaavasti pitää paikkansa, että

$$\mathbf{y} = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{z} + \boldsymbol{\mu} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{4.2.1}$$

missä $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}$ ja $\boldsymbol{\Sigma}^{\frac{1}{2}}$ on $\boldsymbol{\Sigma}$:n symmetrinen neliöjuuri.

4.2.3 χ^2–jakauma

Jos $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I}_p)$, niin

$$\mathbf{z}'\mathbf{z} = z_1^2 + z_2^2 + \ldots + z_p^2 \sim \chi_p^2.$$

Sanomme, että $\mathbf{z}'\mathbf{z}$ noudattaa χ^2–jakaumaa vapausastein p. Jos $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, niin (4.2.1):n avulla on helppo näyttää, että

$$(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \sim \chi_p^2$$

4.2.4 Reunajakaumat

Olkoon $\mathbf{y}_1 = (y_1, y_2, \ldots, y_r)'$ r–ulotteinen ja $\mathbf{y}_2 = (y_{r+1}, y_{r+2}, \ldots, y_p)'$ $(p-r)$–ulotteinen satunnaisvektori.

(a) Jos $\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, missä

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

$\mathrm{E}(\mathbf{y}_1) = \underset{r\times 1}{\boldsymbol{\mu}_1}$, $\mathrm{cov}(\mathbf{y}_1) = \underset{r\times r}{\boldsymbol{\Sigma}_{11}}$, $\mathrm{cov}(\mathbf{y}_1, \mathbf{y}_2) = \boldsymbol{\Sigma}_{12}$, $\mathrm{E}(\mathbf{y}_2) = \boldsymbol{\mu}_2$ ja $\mathrm{cov}(\mathbf{y}_2) = \boldsymbol{\Sigma}_{22}$, niin

$$\mathbf{y}_1 \sim N_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \text{ ja } \mathbf{y}_2 \sim N_{p-r}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}).$$

(b) Erityisesti jokainen $\mathbf{y}$:n alkio y_i, $i = 1, 2, \ldots, p$ noudattaa normaalijakaumaa

$$y_i \sim N(\mu_i, \sigma_{ii}).$$

Olkoon havaintovektori ositettu osavektoreihin $\mathbf{y}$ ja $\mathbf{x}$ siten, että

$$E\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix} \quad \text{ja} \quad \mathrm{cov}\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix}.$$

4.2.5 Riippumattomuus

(a) Jos $\boldsymbol{\Sigma}_{yx} = \mathbf{0}$, niin $\mathbf{y} \perp \mathbf{x}$ (riippumattomia).

(b) Jos $\sigma_{ij} = 0$, niin $y_i \perp y_j$.

4.2.6 Ehdollinen jakauma

Jos

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix} \right],$$

niin $\mathbf{y}|\mathbf{x} = \mathbf{a}$, $\mathbf{y}$:n ehdollinen jakauma ehdolla $\mathbf{x} = \mathbf{a}$, noudattaa multinormaalijakaumaa parametrein

$$\mathrm{E}(\mathbf{y}|\mathbf{x} = \mathbf{a}) = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{a} - \boldsymbol{\mu}_x) \tag{4.2.2}$$

ja

$$\mathrm{cov}(\mathbf{y}|\mathbf{x} = \mathbf{a}) = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}. \tag{4.2.3}$$

4.2.7 Satunnaisvektorien summan jakauma

Olkoot $\mathbf{y}$ ja $\mathbf{x}$ $p \times 1$–vektoreita ja toisistaan riippumattomia. Silloin

$$\begin{aligned} \mathbf{y}+\mathbf{x} &\sim N_p(\boldsymbol{\mu}_y+\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{yy}+\boldsymbol{\Sigma}_{xx}) \\ \mathbf{y}-\mathbf{x} &\sim N_p(\boldsymbol{\mu}_y-\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{yy}+\boldsymbol{\Sigma}_{xx}) \end{aligned}$$

Esimerkki 4.1 Ehdollinen jakauma. Tarkastellaan kaksiulotteista normaalijakaumaa

$$\begin{pmatrix} y \\ x \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{xy} & \sigma_x^2 \end{pmatrix} \right].$$

Olkoon $\boldsymbol{\Sigma} = \mathbf{T}\mathbf{T}'$ kovarianssimatriisin $\boldsymbol{\Sigma}$ Choleskyn hajotelma. Silloin

$$\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} t_{11} & t_{12} \\ 0 & t_{22} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} + \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix},$$

missä $\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \sim N_2(\mathbf{0}, \mathbf{I}_2)$. Saamme yhtälöparin

$$\begin{aligned} y &= t_{11}z_1 + t_{12}z_2 + \mu_y \\ x &= t_{22}z_2 + \mu_x. \end{aligned}$$

Olkoon $x = a$ annettu. Silloin

$$z_2 = t_{22}^{-1}(a - \mu_x)$$

ja

$$y = t_{11}z_1 + t_{12}t_{22}^{-1}(a - \mu_x$$

) + μ_y.Ehdollisensatunnaismuuttujany $|x = a$ jakauman parametrit ovat

$$\mathrm{E}(y|x=a) = \mu_y + t_{12}t_{22}^{-1}(a - \mu_x)$$

ja

$$\mathrm{var}(y|x=a) = t_{11}^2.$$

Choleskyn hajotelman nojalla

$$\begin{pmatrix} t_{11}^2 + t_{12} & t_{12}t_{22} \\ t_{12}t_{22} & t_{22}^2 \end{pmatrix} = \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{xy} & \sigma_x^2 \end{pmatrix},$$

joten

$$\mathrm{E}(y|x=a) = \mu_y + \frac{\sigma_{yx}}{\sigma_x^2}(a - \mu_x)$$

ja

$$\mathrm{var}(y|x=a) = \sigma_y^2 - \frac{\sigma_{yx}^2}{\sigma_x^2}.$$

□

4.3 Multinormaalijakauman parametrien estimointi

Olkoon $\mathbf{y}_1$, $\mathbf{y}_2$, ..., $\mathbf{y}_n$ otos $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:sta, ts.

$$\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n \overset{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Merkintä i.i.d. (independent, identically distributed) tarkoittaa, että satunnaisvektorit $\mathbf{y}_i$ noudattavat samaa jakaumaa ja ovat riippumattomia.

4.3.1 Parametrien $\boldsymbol{\mu}$ ja $\boldsymbol{\Sigma}$ suurimman uskottavuuden estimaatit

Parametrien $\boldsymbol{\mu}$ ja $\boldsymbol{\Sigma}$ suurimman uskottavuuden estimaatit ovat

$$\widehat{\boldsymbol{\mu}} = \bar{\mathbf{y}} \tag{4.3.4}$$

ja

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})' \\ &= \frac{n-1}{n}\mathbf{S}, \end{aligned} \tag{4.3.5}$$

missä $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$.

4.3.2 Multinormaalijakauman uskottavuusfunktio (likelihood–funktio)

Olkoon

$$\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n \overset{\text{i.i.d.}}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Silloin satunnaisvektoreiden $\mathbf{y}_j$ $(j = 1, 2, \ldots, n)$ tiheysfunktio on

$$f(\mathbf{y}_j; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}_j - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_j - \boldsymbol{\mu})}.$$

Otoksen $\mathbf{y}_1$, $\mathbf{y}_2$, ..., $\mathbf{y}_n$ *yhteisjakauman* tiheysfunktio

$$\begin{aligned} f(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{j=1}^{n}\Big\{\frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}\ (\mathbf{y}_j - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_j - \boldsymbol{\mu})}\Big\} \\ &= \frac{1}{(2\pi)^{\frac{np}{2}}|\boldsymbol{\Sigma}|^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{j=1}^{n}\ (\mathbf{y}_j - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_j - \boldsymbol{\mu})} \end{aligned} \tag{4.3.6}$$

4.3.3 Suurimman uskottavuuden estimointi

Suurimman uskottavuuden estimoinnissa sijoitetaan havaintoarvot $\mathbf{y}_1$, $\mathbf{y}_2$, $\ldots$, $\mathbf{y}_n$ yhteisjakauman tiheysfunktioon ja maksimoidaan näin saatu uskottavuusfunktio parametrien suhteen. Arvot $\boldsymbol{\mu} = \widehat{\boldsymbol{\mu}}$ ja $\boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}}$, jotka maksimoivat uskottavuusfunktion, ovat $\boldsymbol{\mu}$:n ja $\boldsymbol{\Sigma}$:n suurimman uskottavuuden estimaatit. Kun havainnot oletetaan annetuiksi, niin otoksen yhteisjakauman tiheysfunktio on parametrien suhteen uskottavuusfunktio ja sitä merkitään

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = f(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Uskottavuusfunktio L on siis parametrien $\boldsymbol{\mu}$ ja $\boldsymbol{\Sigma}$ funktio ja $\mathbf{y}_1$, $\mathbf{y}_2$, $\ldots$, $\mathbf{y}_n$ ovat annettuja havaintoja.

Parametrien $\boldsymbol{\mu}$ ja $\boldsymbol{\Sigma}$ suurimman uskottavuuden estimaatit $\widehat{\boldsymbol{\mu}}$ ja $\widehat{\boldsymbol{\Sigma}}$ toteuttavat yhtälön

$$\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = L(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}).$$

Funktio (4.3.6) saavuttaa maksiminsa, kun eksponentissa oleva lauseke

$$\frac{1}{2} \sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu})'.$$

minimoidaan. Merkitään nyt

$$\Delta^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \tag{4.3.7}$$

ja sijoitetaan $\Delta^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:n lausekkeeseen

$$\mathbf{y}_i - \boldsymbol{\mu} = (\mathbf{y}_i - \bar{\mathbf{y}}) + (\bar{\mathbf{y}} - \boldsymbol{\mu}).$$

Koska $\sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}}) = 0$, niin

$$\Delta^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \bar{\mathbf{y}}) + n(\bar{\mathbf{y}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}).$$

Koska $\boldsymbol{\Sigma}$ ja vastaavasti $\boldsymbol{\Sigma}^{-1}$ ovat positiivisesti definiittejä, niin

$$n(\bar{\mathbf{y}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}) > 0 \tag{4.3.8}$$

kaikilla $\bar{\mathbf{y}} - \boldsymbol{\mu} \neq \mathbf{0}$. Siksi (4.3.8) saavuttaa miniminsä ($= 0$), kun $\boldsymbol{\mu} = \bar{\mathbf{y}}$. Uskottavuusfunktio

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^{np} (|\boldsymbol{\Sigma}|)^{\frac{n}{2}}} e^{-\frac{1}{2}\Delta^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})}$$

riippuu $\boldsymbol{\mu}$:stä vain eksponentin $-\frac{1}{2}\Delta^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ kautta. $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ on eksponentin $-\frac{1}{2}\Delta^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ suhteen kasvava funktio. Kun $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ maksimoidaan $\boldsymbol{\mu}$:n suhteen, haetaan eksponentin $-\frac{1}{2}\Delta^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ maksimikohta, joka on $\Delta^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:n minimikohta. $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ riippuu siis $\boldsymbol{\mu}$:stä vain termin $n(\bar{\mathbf{y}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})$ kautta. Nyt

$$n(\bar{\mathbf{y}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}) > 0, \quad \boldsymbol{\mu} \neq \bar{\mathbf{y}}$$

ja

$$n(\bar{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}) = 0 \Leftrightarrow \boldsymbol{\mu} = \bar{\mathbf{y}}.$$

Voimme siis päätellä, että

$$\max_{\boldsymbol{\mu}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = L(\bar{\mathbf{y}}, \boldsymbol{\Sigma})$$

kaikilla positiivisesti definiiteillä $\boldsymbol{\Sigma}$:n arvoilla. Siksi $\widehat{\boldsymbol{\mu}} = \bar{\mathbf{y}}$ on $\boldsymbol{\mu}$:n suurimman uskottavuuden estimaatti.

Maksimoidaan sitten funktio

$$L(\widehat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{y})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \bar{y})}.$$

matriisin $\boldsymbol{\Sigma}$ suhteen. Voidaan osoittaa, että

$$\max_{\boldsymbol{\Sigma}} L(\widehat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) = L(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}),$$

missä

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})' = \frac{n-1}{n}\mathbf{S}.$$

4.3.4 Estimaattoreiden $\bar{\mathbf{y}}$ ja S jakaumat

Estimaattorit $\bar{\mathbf{y}}$ ja $\mathbf{S}$ ovat parametrien $\boldsymbol{\mu}$ ja $\boldsymbol{\Sigma}$ *tyhjentäviä* otosfunktioista, kun $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ on otos normaalijakaumasta $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

(a) Otos normaalijakaumasta.

1. $\bar{\mathbf{y}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$.
2. $(n-1)\mathbf{S}$ noudattaa Wishartin jakaumaa vapausastein $n-1$.
3. $\bar{\mathbf{y}}$ ja $\mathbf{S}$ ovat riippumattomat.

(b) Kun $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ on otos *ei-normaalisesta* populaatiosta, jonka odotusarvo on $\boldsymbol{\mu}$ ja kovarianssimatriisi $\boldsymbol{\Sigma}$, niin suurilla n:n arvoilla

$$\bar{\mathbf{y}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$$

likimain. Tämä on keskeisen rajaväittämän monimuuttujainen versio.

4.4 Joitakin jakaumatuloksia

4.4.1 Suurten lukujen laki

Suurten lukujen laki. Olkoon $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ otos samasta jakaumasta ($\mathbf{y}_i$:t riippumattomia). Jos $\mathrm{E}(\mathbf{y}_i) = \boldsymbol{\mu}$ on olemassa, silloin kaikilla $\varepsilon > 0$ otoskoon n kasvaessa

$$P(||\bar{\mathbf{y}} - \boldsymbol{\mu}|| > \varepsilon) \to 0.$$

Suurten lukujen lain mukaan $\bar{\mathbf{y}}$ on suurella todennäköisyydellä "lähellä" odotusarvoa $\boldsymbol{\mu}$, kun n on suuri.

4.4.2 Keskeinen rajaväittämä

Olkoon $y_1, y_2, \ldots, y_n$ otos samasta yksiulotteisesta jakaumasta, jonka odotusarvo $\mu = \mathrm{E}(y_i)$ ja $\sigma^2 = \mathrm{var}(y_i)$ ovat olemassa. Silloin jokaista annettua β kohti

$$P\Big(\frac{\bar{y} - \mu}{\sigma/\sqrt{n}} < \beta\Big) \to \Phi(\beta),$$

kun n kasvaa. Funktio $\Phi(\beta) = \int_{-\infty}^{\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz$ on normeeratun normaalijakauman kertymäfunktio.

Monimuuttujaisen keskeisen rajaväittämän mukaan

$$\sqrt{n}(\bar{\mathbf{y}} - \boldsymbol{\mu}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu})$$

noudattaa likimain jakaumaa $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ suurilla n:n arvoilla, kun $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ on otos p–ulotteisesta populaatiosta (normaalijakautunut), jonka odotusarvo $\boldsymbol{\mu} = \mathrm{E}(\mathbf{y}_i)$ ja kovarianssimatriisi $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{y}_i)$ ovat olemassa.

4.4.3 Normaalijakaumaan liittyviä jakaumia

χ^2–jakauma ja Wishart–jakauma ovat normaalijakaumaan läheisesti liittyviä jakaumia.

χ^2–jakauma

Jos $z_i \sim N(0,1)$, $i = 1, 2, \ldots, n$, niin

$$\sum_{i=1}^{n} z_i^2 \sim \chi_n^2.$$

Jos esimerkiksi $y_i \sim N(\mu, \sigma^2)$, niin

$$\sum_{i=1}^{n} \Big(\frac{y_i - \mu}{\sigma}\Big)^2 = \sum_{i=1}^{n} z_i^2 \sim \chi_n^2.$$

Jos μ:n paikalle sijoitetaan $\bar{y}$, niin

$$\sum_{i=1}^{n}\Big(\frac{y_i-\bar{y}}{\sigma}\Big)^2=\frac{(n-1)s^2}{\sigma^2}\sim\chi^2_{n-1}.$$

Wishartin jakauma

Wishartin jakauma on χ^2–jakauman moniulotteinen yleistys:

$$\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})(\mathbf{y}_i-\boldsymbol{\mu})'\sim W_p(n,\boldsymbol{\Sigma}), \tag{4.17}$$

kun $\mathbf{y}_1$, $\mathbf{y}_2$, ..., $\mathbf{y}_n$ on otos $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$:sta. Kun odotusarvo $\boldsymbol{\mu}$ lausekkeessa (4.17) korvataan otoskeskiarvolla $\bar{y}$, noudattaa satunnaismatriisi edelleen Wishartin jakaumaa

$$\sum_{i=1}^{n}(\mathbf{y}_i-\bar{\mathbf{y}})(\mathbf{y}_i-\bar{\mathbf{y}})'\sim W_p(n-1,\boldsymbol{\Sigma}), \tag{4.18}$$

mutta vapausasteet vähenevät yhdellä.

4.5 Jakaumahypoteesin tutkiminen

Multinormaalisuuden toteamiseksi on olemassa monia graafisia menetelmiä ja testejä. Olkoot $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ p–ulotteisia havaintoja. Silloin voidaan tietysti tutkia jokaisen yksittäisen muuttujan y_i normaalisuutta. Vaikka jokainen y_i, $i = 1, 2, \ldots, p$ olisi normaalisti jakautunut, ei p–ulotteinen $\mathbf{y}$ välttämättä ole normaalisti jakautunut. Vaikka 2–ulotteisen ja vielä 3–ulotteisenkin jakauman tuottamia pisteparvia voidaan vielä tutkia havainnollisesti, on moniulotteisten havaintojen jakautuneisuuden tutkiminen yleisesti ottaen vaikeaa.

4.5.1 Yksiulotteinen normaalisuus

Esitämme tässä vain yhden graafisen menettelyn normaalisuuden toteamiseksi, vaikka menetelmiä onkin runsaasti.

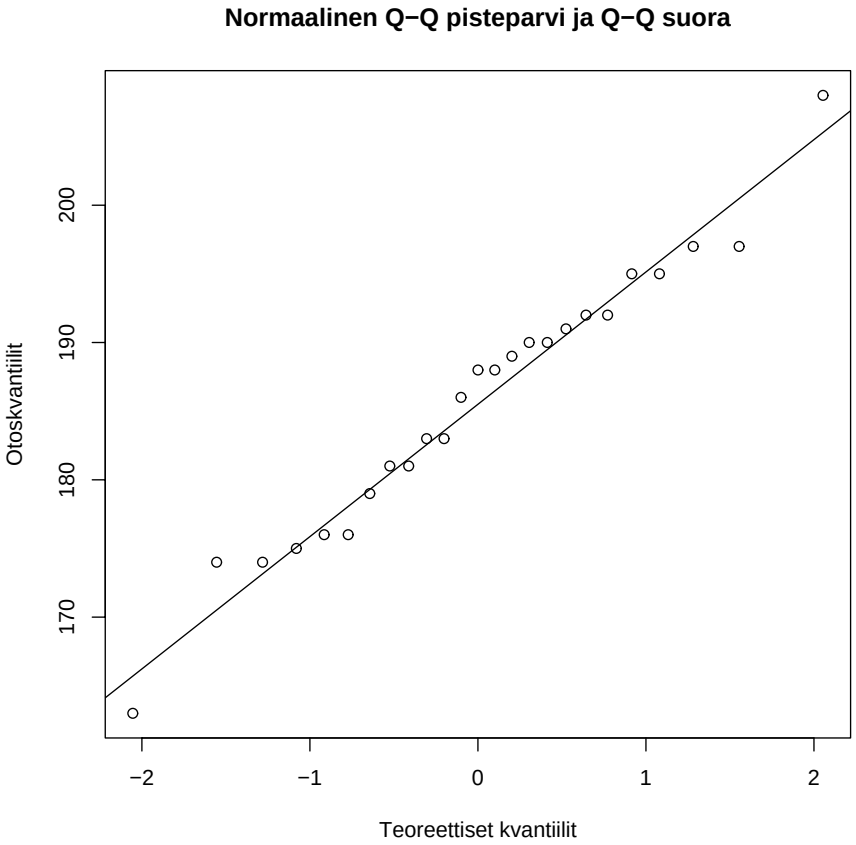
Q–Q–käyrä

Q–Q–käyrällä vertaillaan otoksesta laskettuja kvantiileita vastaaviin populaation kvantiileihin. Otoskvantiilien laskemista varten havainnot $y_1, y_2, \ldots, y_n$ järjestetään ja merkitään järjestettyä otosta $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$, jossa $y_{(1)} \le y_{(2)} \le \ldots \le y_{(n)}$. Järjestyksessä i. havainto $y_{(i)}$ on i/n otoskvantiili. Luvun i/n sijasta käytetään usein lukua $\frac{i-\frac{1}{2}}{n}$ (jatkuvuuskorjaus).

Arvoihin $(i - \frac{1}{2})/n$ liittyvät populaation kvantiilit q_i määritellään standardoidun normaalijakauman avulla seuraavasti

$$\Phi(q_i) = P(y < q_i) = \frac{i - \frac{1}{2}}{n}.$$

Sen jälkeen piirretään pisteparvi $(q_i, y_{(i)})$. Syntyvän Q–Q–käyrän tulisi olla lineaarinen, jos otos on normaalinen.



Kuva 4.1: Pään pituuden `Q-Q pisteparvi` ja `Q-Q suora`, kun on aineistona 25 perheen 1. poikalasten pään pituudet.

4.5.2 Multinormaalisuuden tutkiminen

Multinormaalisuuden tarkistaminen on vaikea tehtävä. Monia menetelmiä on esitetty, mutta täysin tyydyttävää keinoa ei ole tarjolla. Tässä esitetään vain eräs tapa tarkastella multinormaalisuutta. Olkoon $\mathbf{y}$ p–ulotteinen satunnaismuuttuja, jonka odotusarvo on $\boldsymbol{\mu}$ ja kovarianssimatriisi $\boldsymbol{\Sigma}$. Eräs tapa

määritellä multinormaalijakauma on seuraava: Jos $\mathbf{a}'\mathbf{y} \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$ kaikilla $p \times 1$–vektorin $\mathbf{a}$ arvoilla, niin $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Olkoon $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ tarkasteltava otos. Valitaan nyt useita $\mathbf{a}$:n arvoja $\mathbf{a} = \mathbf{a}_\alpha$, $\alpha = 1, 2, \ldots, m$. Kertomalla havainnot $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ suuntavektorilla $\mathbf{a}_\alpha$ saadaan satunnaismuuttujat $v_i = \mathbf{a}'_\alpha \mathbf{y}_i$, $i = 1, 2, \ldots, n$. Näiden v–arvojen tulisi noudattaa 1–ulotteista normaalijakaumaa. Valitaan useita vektoreita $\mathbf{a}_\alpha$ ja tutkitaan v–arvot Q–Q–käyrällä.

Multinormaalisuutta voidaan tarkastell myös Mahalanobisin etäisyyksien

$$D_i^2 = (\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{S}^{-1} (\mathbf{y}_i - \bar{\mathbf{y}}), \quad i = 1, 2, \ldots, n$$

avulla. Voidaan osoittaa, että

$$u_i = \frac{n D_i^2}{(n-1)^2}$$

noudattaa beta-jakaumaa, kun $\mathbf{y}_i$ noudattaa normaalijakaumaa. Arvojen $u_1, u_2, \ldots, u_n$ perusteella voidaan laatia beta-jakaumaan perustuva $Q - Q$ pisteparvi. Pisteiden epälineaarinen ryhmitys viittaisi siihen, että havainnot eivät noudata multinormaalijakaumaa. On myös olemassa maksimiarvoon $D_{(n)}^2 = \max D_i^2$ perustuva normaalisuustesti.

Harjoituksia

1. Olkoon
$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$
missä $\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ ja $\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 1 \\ 1 & 1/2 \end{pmatrix}$.
   (a) Laske $\boldsymbol{\Sigma}$:n Choleskyn hajotelma $\boldsymbol{\Sigma} = \mathbf{T}\mathbf{T}'$, missä $\mathbf{T}$ on yläkolmiomatriisi.
   (b) Kirjoita y_1:n ja y_2:n lausekkeet, kun $\mathbf{y} = \mathbf{T}\mathbf{z} + \boldsymbol{\mu}$ ja $\mathbf{z} \sim N_2(\mathbf{0}, \mathbf{I}_2)$, missä $\mathbf{z}' = (z_1, z_2)$.

2. (Jatkoa edelliseen) Kirjoita auki $\mathbf{y}$:n tiheysfunktio $\mathbf{z}$:n avulla, kun $\mathbf{z} = \mathbf{T}^{-1}(\mathbf{y} - \boldsymbol{\mu})$.

3. Oletetaan, että $\mathbf{y} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, missä $\boldsymbol{\mu} = (\,1 \quad 1 \quad 2\,)'$ ja
$$\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 6 & 10 \\ 6 & 58 & 29 \\ 10 & 29 & 38 \end{pmatrix}.$$
Olkoon $\mathbf{y}_2 = \begin{pmatrix} y_2 \\ y_3 \end{pmatrix}$. Laske
(a) $\mathrm{E}(\mathbf{y}_2 | y_1 = 2)$
(b) $\mathrm{cov}(\mathbf{y}_2 | y_1 = 2)$.

4. Simuloi 10, 30 ja 100 alkion otokset normaalijakaumasta $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, missä parametrien $\boldsymbol{\mu}$ ja $\boldsymbol{\Sigma}$ arvot ovat samat kuin tehtävässä 3. Laske otoskeskiarvot ja otoskovarianssit.

5. Oletetaan, että $\mathbf{y}_1 = \begin{pmatrix} 3 \\ 6 \end{pmatrix}, \mathbf{y}_2 = \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \mathbf{y}_3 = \begin{pmatrix} 5 \\ 7 \end{pmatrix}, \mathbf{y}_4 = \begin{pmatrix} 4 \\ 7 \end{pmatrix}$ on otos normaalijakaumasta $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Määrää parametrien $\boldsymbol{\mu}$ ja $\boldsymbol{\Sigma}$ suurimman uskottavuuden estimaatit.

6. Olkoon $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{20}$ otos normaalijakaumasta $N_6(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Määrittele seuraavat jakaumat:
   (a) Satunnaismuuttujan $(\mathbf{y}_1 - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu})$ jakauma.
   (b) Satunnaismuuttujan $\bar{\mathbf{y}}$ jakauma.
   (c) Satunnaismuuttujan $\sqrt{n}(\bar{\mathbf{y}} - \boldsymbol{\mu})$ jakauma.

7. Sijoita suurimman uskottavuuden estimaatit $\hat{\boldsymbol{\mu}}$ ja $\hat{\boldsymbol{\Sigma}}$ uskottavuusfunktioon ja näytä, että p-ulotteisen normaalijakauman uskottavuusfunktion maksimi on

$$L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \frac{1}{(2\pi)^{np/2}} e^{-np/2} \frac{1}{|\hat{\boldsymbol{\Sigma}}|^{n/2}}.$$

8. Satunnaisvektorit $\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{y}_3$ ja $\mathbf{y}_4$ ovat riippumattomat ja noudattavat normaalijakaumaa $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
   (a) Mitä ovat satunnaisvektorien

$$\mathbf{v}_1 = \frac{1}{4}\mathbf{y}_1 - \frac{1}{4}\mathbf{y}_2 + \frac{1}{4}\mathbf{y}_3 - \frac{1}{4}\mathbf{y}_4 \text{ ja } \mathbf{v}_2 = \frac{1}{4}\mathbf{y}_1 + \frac{1}{4}\mathbf{y}_2 - \frac{1}{4}\mathbf{y}_3 - \frac{1}{4}\mathbf{y}_4$$

   jakaumat?
   (b) Määritä satunnaisvektorin

$$\mathbf{v} = \begin{pmatrix} \mathbf{v}_1' \\ \mathbf{v}_2' \end{pmatrix}$$

   jakauma.

9. Tarkastellaan kahden muuttujan y_1, y_2 normaalijakumaa, jonka odotusarvo $\mathrm{E}(y_1) = \mu_1 = 0$, $\mathrm{E}(y_2) = \mu_2 = 2$, $\mathrm{var}(y_1) = \sigma_{11} = 2$, $\mathrm{var}(y_2) = \sigma_{22} = 1$ ja muuttujien välinen korrelaatio $\rho_{12} = 0.5$.
   (a) Kirjoita kyseinen kahden muuttujan normaalijakauman tiheysfunktio.
   (b) Mikä on funktion $(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = 1$ kuvaaja? (y_1 ja y_2 reaalimuuttujia)1

10. Satunnaisvektori $\mathbf{y}$ noudattaa normaalijakaumaa $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ parametrein
$$\boldsymbol{\mu} = \begin{pmatrix} -3 \\ 1 \\ 4 \end{pmatrix} \text{ ja } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$
Mitkä seuraavista satunnaismuuttujista ovat riippumattomat?
(a) y_1 ja y_2,
(b) y_1 ja y_3,
(c) $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ ja y_3,
(d) $\frac{1}{2}(y_1 + y_3)$ ja y_3,
(e) y_2 ja $y_2 - \frac{5}{2}y_1 - y_3$.

11. Olkoon jakauma sama kuin tehtävässä 9. Määritä
(a) y_1:n ehdollinen jakauma, kun $y_2 = a$ on annettu.
(b) y_2:n ehdollinen jakauma, kun $y_1 = a$ ja $y_3 = b$.

12. Tutki `Pojat`aineistosta, noudattavatko 1. poikien pään mitat 2-ulotteista normaalijakaumaa.

    (a) Piirrä $Q - Q$-pisteparvet ja suorat yksittäisistä muuttujista.

    (b) Tarkastele samalla tavalla pääkomponenttien jakaumia.

    (c) Tutki jakautuneisuutta myös havaintojen Mahalanobisin etäisyyksien D_i avulla.

13. Oletetaan, että $\mathbf{y} \sim \mathrm{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, missä
$$\boldsymbol{\mu} = \begin{pmatrix} 3 \\ 1 \\ 4 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 6 & 1 & -2 \\ 1 & 13 & 4 \\ -2 & 4 & 13 \end{pmatrix}.$$

    (a) Määritä satunnaismuuttujan $z = 2y_1 - y_2 + 3y_3$ jakauma ja

    (b) satunnaismuuttujien $z_1 = y_1 + y_2 + y_3$ ja $z_2 = y_1 - y_2 + 2y_3$ yhteisjakauma.

    (c) Mikä on y_2:n jakauma,

    (d) y_1:n ja y_3:n yhteisjakauma?

    (e) y_1:n,y_3:n ja $\frac{1}{2}(y_1 + y_2)$:n yhteisjakauma?

Luku 5

Monimuuttujaisia testejä

5.1 Monimuuttujaiset vai yhden muuttujan testit

Olkoon $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ p–ulotteinen otos populaatiosta, jonka odotusarvo on $\boldsymbol{\mu}$ ja kovarianssimatriisi $\boldsymbol{\Sigma}$. Testataan esimerkiksi hypoteesia

$$H_0: \quad \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{vs.} \quad H_1: \quad \boldsymbol{\mu} \neq \boldsymbol{\mu}_0,$$

missä $\boldsymbol{\mu}_0 = \begin{pmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \mu_{0p} \end{pmatrix}$ on jokin vektorin $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$ annettu arvo. Seuraavassa yleistetään tutut keskiarvotestit moniulotteiseen tapaukseen.

5.2 Keskiarvon $\boldsymbol{\mu}$ testaus, kun $\boldsymbol{\Sigma}$ tunnetaan

5.2.1 Yhden muuttujan testin $H_0: \mu = \mu_0$

Testataan

$$H_0: \quad \mu = \mu_0 \quad \text{vs.} \quad H_1: \quad \mu \neq \mu_0.$$

Olkoon $y_1, y_2, \ldots, y_n \sim N(\mu, \sigma^2)$, missä σ^2 tunnetaan. Silloin

$$z = \frac{\bar{y} - \mu_0}{\sigma_{\bar{y}}} = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1), \tag{5.2.1}$$

jos H_0 on tosi. Vastaavasti

$$z^2 \sim \Big(\frac{\bar{y} - \mu}{\sigma/\sqrt{n}}\Big)^2 = n\Big(\frac{\bar{y} - \mu}{\sigma}\Big) = n\boldsymbol{\Delta}^2 \sim \chi^2_1.$$

Vaikka havainnot *eivät olisikaan peräisin normaalijakaumasta*, niin keskeisen rajaväittämän mukaan

$$\frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$$

noudattaa suurilla n:n arvoilla likimain normaalijakaumaa $\mathrm{N}(0,1)$.

5.2.2 Monimuuttujainen testi $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$, kun $\boldsymbol{\Sigma}$ tunnetaan

Olkoon $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ otos $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:sta, missä $\boldsymbol{\Sigma}$ tunnetaan. Testataan hypoteesi

$$H_0: \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} \mu_{01} \\ \mu_{02} \\ \vdots \\ \mu_{0p} \end{pmatrix} \quad \text{vs.} \quad H_1: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0,$$

missä μ_{0i}:t ovat annettuja. Jos H_0 on tosi, niin $\bar{\mathbf{y}} \sim N_p(\mu_0, \frac{1}{n}\boldsymbol{\Sigma})$ ja

$$\begin{aligned} z^2 &= (\bar{\mathbf{y}} - \boldsymbol{\mu}_0)' \Big(\frac{1}{n}\boldsymbol{\Sigma}\Big)^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_0) \\ &= n\boldsymbol{\Delta}^2 = n(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_0) \sim \chi^2_p. \end{aligned} \tag{5.2.2}$$

Jos $\boldsymbol{\Sigma}$ on tuntematon, niin $\boldsymbol{\Sigma}$ voidaan korvata $\mathbf{S}$:llä lausekkeessa (5.2.2). Silloin silloin testisuure ei enää noudata χ^2-jakaumaa. Tosin $n(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_0)$ noudattaa likimain jakaumaa χ^2-jakaumaa vapausastein p, kun n on suuri verrattuna p:n arvoon.

5.3 Keskiarvotesti, kun Σ tuntematon

5.3.1 Yhden muuttujan t–testi

Olkoon $y_1, y_2, \ldots, y_n$ otos $N(\mu, \sigma^2)$, missä σ on tuntematon. Käytetään μ:n ja σ^2:n estimaattoreita $\bar{y}$ ja s^2. Testin

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu = \mu_0$$

testisuureena käytetään

$$t = \frac{y - \mu_0}{s/\sqrt{n}} = \frac{\sqrt{n}(y - \mu_0)}{s}. \tag{5.3.3}$$

Jos H_0 on tosi, niin

$$t \sim t_{n-1}$$

eli t noudattaa t-jakaumaa vapausastein $n - 1$.

5.3.2 Hotellingin T–testi, kun Σ tuntematon ja $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$

Olkoon $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ otos $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:sta, missä $\boldsymbol{\mu}$ ja $\boldsymbol{\Sigma}$ tuntemattomia. Parametrien $\boldsymbol{\mu}$ ja $\boldsymbol{\Sigma}$ estimaatit ovat $\bar{\mathbf{y}}$ ja $\mathbf{S}$. Yhden muuttuja testisuure (5.3.3) voidaan kirjoittaa muodossa

$$t^2 = \frac{n(\bar{y} - \mu_0)^2}{s^2} = n(\bar{y} - \mu_0)(s^2)^{-1}(\bar{y} - \mu_0).$$

Voidaan osoittaa, että moniulotteisessa tapauksessa t-testisuureen vastine on

$$T^2 = n(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}_0). \tag{5.3.4}$$

On helppo huomata, että T^2 saadaan z^2:sta korvaamalla $\boldsymbol{\Sigma}$ estimaatillaan $\mathbf{S}$. Pelkästään tämä huomio ei tietenkään ole perustelu sille, että T^2 olisi hyvä testisuure. *Hotelling* (1931) johti T^2:n jakauman, jota kutsutaan hänen mukaansa Hotellingin T^2-jakaumaksi. T^2–jakauman parametrit ovat p ja $n-1$ ja merkitään $T^2 \sim T^2_{p,n-1}$. Nollahypoteesi $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ (vs. $H_1 : \ \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$) hylätään, jos $T^2 > T_{\alpha,p,n-1}$ eli T^2:n arvo ylittää T^2–jakauman kriittisen arvon riskitasolla α.

Jos $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, niin $z = \mathbf{a}'\mathbf{y} \sim N_p(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$. Silloin muuttujan z:n otoskeskiarvo ja varianssi ovat

$$\bar{\mathbf{z}} = \mathbf{a}'\bar{\mathbf{y}} \qquad \text{ja} \qquad s_z^2 = \mathbf{a}'\mathbf{S}\mathbf{a},$$

missä $\bar{\mathbf{y}}$ ja $\mathbf{S}$ ovat moniulotteisten parametrien μ ja $\boldsymbol{\Sigma}$ estimaatit. Nollahypoteesin $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ vallitessa $\mathrm{E}(z) = \mathbf{a}'\boldsymbol{\mu}_0$ ja t-testisuure hypoteesin $\mathbf{a}'\boldsymbol{\mu} = \mathbf{a}'\boldsymbol{\mu}_0$ testaamiseksi on

$$t(\mathbf{a}) = \frac{\mathbf{a}'(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)}{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}/n}}. \tag{5.3.5}$$

Tavallisessa 2-puolisessa t-testissä nollahypoteesi $\mathbf{a}'\boldsymbol{\mu} = \mathbf{a}'\boldsymbol{\mu}_0$ hylätään riskitasolla α, kun

$$\frac{n[\mathbf{a}'(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)]^2}{\mathbf{a}'\mathbf{S}\mathbf{a}} > t^2_{\alpha/2}(n-1),$$

missä $t_{\alpha/2}(n-1)$ on $t(n-1)$-jakauman $(1-\alpha/2)100\%$:n piste.

Jos nollahypoteesi $\mathbf{a}'\boldsymbol{\mu} = \mathbf{a}'\boldsymbol{\mu}_0$ hylätään, niin otosinformaatio on ristiriidassa oletuksen $\mathrm{E}(z) = \mathbf{a}'\boldsymbol{\mu}_0$ kanssa. Koska $\mathbf{a}$ on annettu vakiovektori, seuraa hypoteesin $\mathbf{a}'\boldsymbol{\mu} = \mathbf{a}'\boldsymbol{\mu}_0$ hylkäämisestä, että myös hypoteesi $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ on hylättävä. Jos oletetaan esimerkiksi, että

$$\boldsymbol{\mu} = \begin{pmatrix} 3 \\ 1 \\ 5 \end{pmatrix} \qquad \text{ja} \qquad \boldsymbol{\mu}_0 = \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix}.$$

Silloin moniulotteinen hypoteesi $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ ei pidä paikaansa. Jos $\mathbf{a}$ on muotoa $\mathbf{a} = (a_1, a_2, 0)$, niin kaikki yksiulotteiset hypoteesit $\mathbf{a}'\boldsymbol{\mu} = \mathbf{a}'\boldsymbol{\mu}_0$ pitävät aina paikkansa. Jos sen sijaan $\mathbf{a} = (a_1, a_2, a_3)$, missä $a_3 \neq 0$, niin silloin hypoteesi $\mathbf{a}'\boldsymbol{\mu} = \mathbf{a}'\boldsymbol{\mu}_0$ ei pidä paikkaansa. On siis selvää, että testin tulos riippuu ratkaisevasti myös kerroivektorin $\mathbf{a}$ valinnasta.

Eräs ilmeinen ratkaisu on valita yhtälössä (5.3.5) $\mathbf{a}$ siten, että $t(\mathbf{a})$ saa mahdollisimman suuren arvon. Siinä tapauksessa otosinformaation ja hypoteesin poikkeama on mahdollisimman suuri. Merkitään näin saatavaa testisuusetta T^2, joten

$$T^2 = \max_{\|\mathbf{a}\|} \frac{n[\mathbf{a}'(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)]^2}{\mathbf{a}'\mathbf{S}\mathbf{a}}. \tag{5.3.6}$$

Vektorin $\mathbf{a}$ pituus ei vaikuta T^2:n arvoon, ainoastaan sen suunta. Voidaan osoittaa, että maksimi saavutetaan, kun $\mathbf{a} \propto \mathbf{S}^{-1}(\bar{y} - \boldsymbol{\mu}_0)$. Silloin

$$T^2 = n(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}_0),$$

kuten yhtälössä (5.3.4). Testisuure (5.3.4) voidaan johtaa myös uskottavuussuhdetestinä.

Kun $n > p$, niin

$$T^2 = n(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)$$

noudattaa Hotellingin T^2-jakaumaa $T^2_{p,n-1}(\delta^2)$, missä

$$\delta^2 = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0).$$

Nollahypoteesin $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ vallitessa $\delta^2 = 0$, joten silloin

$$T^2 \sim T^2_{p,n-1}.$$

Jos T^2 noudattaa T^2-jakaumaa vapausastein p ja f, eli

$$T^2 \sim T^2_{p,f},$$

niin silloin

$$\frac{(f-p+1)}{pf}T^2 \sim F(p, f-p+1), \tag{5.3.7}$$

eli $\frac{(f-p+1)}{pf}T^2$ noudattaa F-jakaumaa vapausastein p ja $f-p+1$. Tarkasteltavan T^2-jakautuneen satunnaisvektorin pituus määrittää p:n arvon ja f riippuu havaintojen ja estimoitavien parametrien lukumäärästä. Näin siis Hotellingin T^2-testi voidaan tehdä F-jakauman avulla. Hypoteesia $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ testattaessa $f = n-1$, joten silloin

$$\frac{(n-p)}{p(n-1)}T^2 \sim F(p, n-p). \tag{5.3.8}$$

5.3.3 Luottamusvälit

Jos oletetaan, että $\boldsymbol{\mu}$ on havaintojen oikea (tuntematon) odotusarvo, niin (**??**):n mukaan

$$P[\frac{(n-p)T^2}{p(n-1)} < F_\alpha(p, n-p)] = 1-\alpha, \tag{5.3.9}$$

josta seuraa

$$P[n(\bar{\mathbf{y}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}) < \frac{p(n-1)}{(n-p)}F_\alpha(p, n-p)] = 1-\alpha. \tag{5.3.10}$$

Havaituilla $\bar{\mathbf{y}}$:n ja $\mathbf{S}^{-1}$:n arvoilla voidaan hakasulkeissa oleva epäyhtälö kirjoittaa $\boldsymbol{\mu}$:n funktiona muodossa

$$n(\bar{\mathbf{y}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}) < \frac{p(n-1)}{(n-p)}F_\alpha(p, n-p). \tag{5.3.11}$$

Epäyhtälön (5.3.11) määrittelemä alue on $\boldsymbol{\mu}$:n $100(1-\alpha)$ prosentin luottamusalue. Alue on hyperellipsoidi, jonka keskipiste on pisteessä $\boldsymbol{\mu} = \bar{\mathbf{y}}$. Tällaista moniulotteista aluetta saattaa olla vaikea tarkastella, kun $p > 2$.

Vaihtoehtoinen tarkastelu perustuu siihen havaintoon, että $t(\mathbf{a})$:n maksimi (vrt. yhtälö (5.3.6) noudattaa T^2:jakumaa. Silloin yhtälön (5.3.10) mukaan

$$P[\max_{\|\mathbf{a}\|} \frac{n[\mathbf{a}'(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)]^2}{\mathbf{a}'\mathbf{S}\mathbf{a}} < \frac{p(n-1)}{(n-p)}F_\alpha(p, n-p)] = 1 - \alpha. \tag{5.3.12}$$

Siksi yhtälö

$$P[[\mathbf{a}'(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)]^2 < \frac{p(n-1)}{(n-p)}F_\alpha(p, n-p)\frac{s_a^2}{n}] = 1 - \alpha,$$

missä $s_a^2 = \mathbf{a}'\mathbf{S}\mathbf{a}$, pitää paikkansa kaikilla vektorin $\mathbf{a}$ valinnoilla. Tapahtumat

$$[\mathbf{a}'(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)]^2 < \frac{p(n-1)}{(n-p)}F_\alpha(p, n-p)\frac{s_a^2}{n}$$

ja

$$|\mathbf{a}'(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)| < [\frac{p(n-1)}{(n-p)}F_\alpha(p, n-p)]^{1/2}\frac{s_a}{\sqrt{n}}$$

ovat yhtäpitävät, ja jälkimmäinen tapahtuma voidaan kirjoittaa muodossa

$$\mathbf{a}'\bar{\mathbf{y}} - K_{\alpha/2}\frac{s_a}{\sqrt{n}} < \mathbf{a}'\boldsymbol{\mu} < \mathbf{a}'\bar{\mathbf{y}} + K_{\alpha/2}\frac{s_a}{\sqrt{n}},$$

missä $K_{\alpha/2} = [\frac{p(n-1)}{(n-p)}F_\alpha(p, n-p)]^{1/2}$. Siksi jokaista havaittua $\bar{\mathbf{y}}$ ja $\mathbf{S}$ kohti,

$$\mathbf{a}'\bar{\mathbf{y}} \pm K_{\alpha/2}\frac{s_a}{\sqrt{n}} \qquad \text{kaikilla } \mathbf{a} \tag{5.3.13}$$

on $\mathbf{a}'\boldsymbol{\mu}$:n samanaikaisten luottamusvälien joukko, kun luottamustaso on $100(1-\alpha)\%$.

5.4 Riippuvat otokset

5.4.1 Yhden muuttujan tapaus

Monissa sovelluksissa otokset eivät ole riippumattomia, vaan kahden eri otoksen havainnoilla on tietty luonnollinen vastaavuus. Esimerkiksi tilastoyksiköihin sovelletaan kaksi eri käsittelyä ja i. tilastoyksikön tulos 1. käsittelyssä

on y_i ja 2. käsittelyssä x_i. Silloin mittaluvuilla on luonnollinen vastinpari. Alkuperäinen kahden otoksen tilanne voidaan palauttaa yhteen otokseen siitymällä tarkastelemaan erotuksia. Asetelma olisi siis seuraava:

Taulukko 5.1 Riippuvien otosten testiasetelma.

Pari	Käsittely 1	Käsittely 2	$d_i = y_i - x_i$
1	y_1	x_1	d_1
2	y_2	x_2	d_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n	d_n

Saadaksemme t–testin tarvitsemme tietoa y:n ja x:n yhteisjakaumasta. Oletetaan, että

$$\begin{pmatrix} y \\ x \end{pmatrix} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

missä $\boldsymbol{\mu} = (\mu_y, \mu_x)'$ ja $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{pmatrix}$. Silloin $d_i = y_i - x_i \sim N(\mu_y - \mu_x, \sigma_d^2)$, missä $\sigma_d^2 = \sigma_y^2 + \sigma_x^2 - 2\sigma_{yx}$. Lasketaan otoksesta

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{ja} \quad s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2.$$

Hypoteesin $H_0:\ \mu_y = \mu_x$ (tai $\mu_d = 0$) testaamiseksi käytetään suuretta

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \tag{5.4.14}$$

joka on t_{n-1}, mikäli H_0 on tosi.

5.4.2 Monimuuttujaiset riippuvat otokset

Asetelma on sama kuin yhden muuttujan tapauksessa, mutta havainnot ovat p:n muuttujan vektoreita. Olkoon $\mathbf{y}_i$ ensimmäisestä otoksesta ja $\mathbf{x}_i$ toisesta, $i = 1, 2, \ldots, p$. Silloin meillä on seuraava asetelma:

Taulukko 5.2 Riippuvien otosten monimuuttujainen testiasetelma.

Pari	Käsittely 1	Käsittely 2	$\mathbf{d}_i = \mathbf{y}_i - \mathbf{x}_i$
1	$\mathbf{y}_1$	$\mathbf{x}_1$	$\mathbf{d}_1$
2	$\mathbf{y}_2$	$\mathbf{x}_2$	$\mathbf{d}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$\mathbf{y}_n$	$\mathbf{x}_n$	$\mathbf{d}_n$

Hypoteesin $H_0: \boldsymbol{\mu}_d = \mathbf{0}$ testaamiseksi lasketaan

$$\bar{\mathbf{d}} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{ja} \quad \mathbf{S}_d = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{d}_i - \bar{\mathbf{d}})(\mathbf{d}_i - \bar{\mathbf{d}}).$$

Saadaan testisuure

$$T^2 = \bar{\mathbf{d}}'\Big(\frac{\mathbf{S}_d}{n}\Big)^{-1}\bar{\mathbf{d}} = nD^2, \tag{5.4.15}$$

missä $D^2 = \bar{\mathbf{d}}'\mathbf{S}_d^{-1}\bar{\mathbf{d}}$. H_0 hylätään, kun $T^2 > T^2_{\alpha,p,n-1}$. Jos otokset ovat riippuvia, pitäisi käyttää testiä (5.4.15). Riippumattomien otosten T^2 testi ei ole tässä tapauksessa yhtä voimakas.

5.5 Keskiarvon rakenteen testaaminen

Pykälässä 5.3.2 tarkasteltiin hypoteeseja, jotka spesifioivat keskiarvon täydellisesti. Samalla tekniikalla on mahdollista testata hypoteeseja, jotka spesifioivat keskiarvovektorin komponenttien välille joitakin relaatioita.

Tarkastellaan esimerkiksi tilannetta, jossa jokaselta yksiköltä mitataan sama muuttuja eri olosuhteissa. Tehdään esimerkiksi jokin psykologinen testi eri ajankohtina tai mitataan jonkin suorituksen tulos vuorokauden eri aikoina, sanokaamme p:nä eri ajankohtana. Eräs kiinnostava hypoteesi voisi olla $H_0: \mu_1 = \mu_2 \ldots = \mu_p$. Hypoteesi siis väittää, että kaikkien muuttujien keskiarvot ovat yhtä suuret. Hypoteesi voidaan myös esittää esimerkiksi muodossa $H_0: \mu_1 = \mu_j,\ j = 2, \ldots = \mu_p$. Matriisimuodossa

$$H_0: \mathbf{C}\boldsymbol{\mu} = \mathbf{0},$$

missä

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & \ldots & 0 \\ 1 & 0 & -1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \ldots & -1 \end{pmatrix}$$

on $(p-1) \times p$-matriisi. Nollahypoteesin vallitessa

$$\mathbf{C}\mathbf{y} \sim N_{p-1}(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'),$$

joten H_0 voidaan testata Hotellingin T^2-testisuureella

$$T^2 = n\bar{\mathbf{y}}'\mathbf{C}'(\mathbf{C}\mathbf{S}\mathbf{C}')^{-1}\mathbf{C}\bar{\mathbf{y}}, \tag{5.5.16}$$

missä täytyy olla $n > p-1$. Kun nollahypoteesi on tosi, niin $T^2 \sim T^2_{p-1}(n-1)$, joten

$$F = \frac{n-p+1}{(p-1)(n-1)}T^2 \sim F(p-1, n-p+1).$$

5.6 Kahden keskiarvon vertailu

5.6.1 Kahden otoksen t–testi

Olkoon $y_{11}, y_{12}, \ldots, y_{1n_1}$ otos $N(\mu_1, \sigma_1^2)$ ja $y_{21}, y_{22}, \ldots, y_{2n_2}$ otos $N(\mu_2, \sigma_2^2)$:sta. Oletetaan, että otokset ovat riippumattomia ja $\sigma_1^2 = \sigma_2^2 = \sigma^2$ on tuntematon. Lasketaan otoksista $\bar{y}_1$, $\bar{y}_2$ ja neliösummat

$$\begin{aligned} SS_1 &= \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 = (n_1 - 1)s_1^2 \quad \text{ja} \\ SS_2 &= \sum_{j=2}^{n_2}(y_{2j} - \bar{y}_2) = (n_2 - 1)s_2^2. \end{aligned}$$

Yhdistetty (pooled) varianssi

$$s_{pl}^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

s_{pl}^2 on σ^2:n harhaton estimaattori.

Testauksessa

$$H_0 : \ \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \ \mu_1 \neq \mu_2$$

testisuure

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_{pl}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}, \tag{5.6.17}$$

kun H_0 on tosi. H_0 hylätään, kun $|t| \geq t_{\frac{\alpha}{2},n_1+n_2-2}$. Huomaa, että $s_{pl}\sqrt{1/n_1 + 1/n_2}$ on hajonnan $\sigma_{\bar{y}_1-\bar{y}_2}$ estimaatti.

5.6.2 Monimuuttujainen 2–otoksen T^2–testi

Testataan hypoteesi

$$H_0 : \ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad H_1 : \ \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2.$$

Olkoon $\mathbf{y}_{11}, \mathbf{y}_{12}, \ldots, \mathbf{y}_{1n_1} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ ja $\mathbf{y}_{21}, \mathbf{y}_{22}, \ldots, \mathbf{y}_{2n_2} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Oletetaan, että otokset ovat toisistaan riippumattomat ja $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ on tuntematon. Lasketaan $\bar{\mathbf{y}}_1 = \sum_{j=1}^{n_1} \mathbf{y}_{1j}/n_1$, $\bar{\mathbf{y}}_2 = \sum_{j=1}^{n_2} \mathbf{y}_{2j}/n_2$ sekä neliösumma-matriisit

$$\begin{aligned} \mathbf{W}_1 &= \sum_{j=1}^{n_1}(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1)(\bar{\mathbf{y}}_{1j} - \bar{\mathbf{y}}_1)' = (n_1 - 1)\mathbf{S}_1 \\ \mathbf{W}_2 &= \sum_{j=1}^{n_2}(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2)(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2)' = (n_2 - 1)\mathbf{S}_2. \end{aligned}$$

$\mathbf{\Sigma}$:n harhaton yhdistetty estimaattori on Matriisi

$$\begin{aligned}\mathbf{S}_{pl} &= \frac{1}{n_1+n_2-2}(\mathbf{W}_1+\mathbf{W}_2)\\ &= \frac{1}{n_1+n_2-2}[(n_1-1)\mathbf{S}_1+(n_2-1)\mathbf{S}_2].\end{aligned}$$

on $\mathbf{\Sigma}$:n harhaton estimaattori. Sitä sanotaan myös yhdistetyksi estimattoriksi, koska se on muodostettu yhdistämällä kahden eri otoksen informaatio.

Yhden muuttujan tapauksessa

$$t^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{y}_1-\bar{y}_2)(s_{pl}^2)^{-1}(\bar{y}_1-\bar{y}_2).$$

Tämä voidaan jälleen yleistää p:n muuttujan tapaukseen sijoittamalla yhden muuttujan testisuureeseen $(\bar{y}_1-\bar{y}_2)$:n paikalle $\bar{\mathbf{y}}_1-\bar{\mathbf{y}}_2$ ja s_{pl}^2:n paikalle $\mathbf{S}_{pl}$. Silloin saadaan testisuure

$$T^2 = \frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{y}}_1-\bar{\mathbf{y}}_2)'\mathbf{S}_{pl}^{-1}(\mathbf{y}_1-\bar{\mathbf{y}}_2) \tag{5.6.18}$$

Testisuure noudattaa T^2–jakaumaa vapausastein p ja n_1+n_2-2, kun H_0 : $\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$ on tosi. Hypoteesi voidaan jälleen testata F-testillä, sillä

$$\frac{n_1+n_2-p-1}{p(n_1+n_2-2)}T^2 \sim F(p, n_1+n_2-p-1). \tag{5.6.19}$$

5.6.3 Uskottavuussuhteeseen perustuva testaus

Olkoon $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ otos $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$:sta. Uskottavuusfunktio saadaan otoksen yhteisjakauman tiheysfunktiosta, kun sitä tarkastellaan parametrien funktiona.

$$\begin{aligned}L(\boldsymbol{\mu}, \mathbf{\Sigma}; \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) &= f(\mathbf{y}_1; \boldsymbol{\mu}, \mathbf{\Sigma}) f(\mathbf{y}_2; \boldsymbol{\mu}, \mathbf{\Sigma}) \cdot \ldots \cdot f(\mathbf{y}_n; \boldsymbol{\mu}, \mathbf{\Sigma})\\ &= \frac{1}{(\sqrt{2\pi})^{np}\left(\sqrt{|\mathbf{\Sigma}|}\right)^n} e^{-\frac{1}{2}\sum_{j=1}^{n}(\mathbf{y}_j-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{y}_j-\boldsymbol{\mu})'}\end{aligned} \tag{5.6.20}$$

missä siis

$$f(\mathbf{y}_j; \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{(\sqrt{2\pi})^p\sqrt{|\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{y}_j-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{y}_j-\boldsymbol{\mu})'}.$$

Olemme jo aikaisemmin todenneet, että $\boldsymbol{\mu}$:n ja $\mathbf{\Sigma}$:n suurimman uskottavuuden estimaatit ovat $\widehat{\boldsymbol{\mu}} = \bar{\mathbf{y}}$ ja $\widehat{\mathbf{\Sigma}} = \frac{n-1}{n}\mathbf{S}$. Sijoittamalla nämä arvot uskottavuusfunktion lausekkeeseen saadaan funktion maksimiarvo

$$\max_{\boldsymbol{\mu}, \mathbf{\Sigma}} L(\boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{np}{2}}|\widehat{\mathbf{\Sigma}}|^{n/2}} e^{-\frac{np}{2}}$$

$$= \quad \text{vakio} \times |\widehat{\boldsymbol{\Sigma}}|^{-\frac{n}{2}}.$$

Hypoteesin

$$H_0: \quad \boldsymbol{\mu} = \boldsymbol{\mu}_0$$

vallitessa

$$L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^{\frac{np}{2}} |\boldsymbol{\Sigma}|^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{j=1}^{n} (\mathbf{y}_j - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_j - \boldsymbol{\mu}_0)}.$$

Koska keskiarvo on kiinnitetty arvoon $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, niin

$$\max_{\boldsymbol{\Sigma}} L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{np}{2}} |\widehat{\boldsymbol{\Sigma}}_0|^{\frac{n}{2}}} e^{-\frac{np}{2}},$$

missä

$$\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{n} \sum_{j=1}^{n} (\mathbf{y}_j - \boldsymbol{\mu}_0)(\mathbf{y}_j - \boldsymbol{\mu}_0)'.$$

Uskottavuussuhde on silloin

$$\Lambda = \frac{\max_{\boldsymbol{\Sigma}} L(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})}{\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})} = \left(\frac{|\widehat{\boldsymbol{\Sigma}}|}{|\widehat{\boldsymbol{\Sigma}}_0|} \right)^{\frac{n}{2}}.$$

Wilksin lambda

$$\Lambda^{\frac{2}{n}} = \frac{|\widehat{\boldsymbol{\Sigma}}|}{|\widehat{\boldsymbol{\Sigma}}_0|}$$

uskottavuussuhteen kanssa yhtäpitävä testisuure. Voidaan myös osoittaa, että

$$\Lambda^{\frac{2}{n}} = \left(1 + \frac{T^2}{n-1}\right)^{-1},$$

missä $T^2 = n(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_0)$ on *Hotellingin* T^2. Uskottavuussuhteeseen perustuva lähestymistapa johtaa Hotellingin T^2–testiin myös kahden otoksen tapauksessa testattaessa hypoteesia $H_0: \ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

Uskottavuussuhteeseen perustuva teniikka tarjoaa yleisen menetelmän testisuureen johtamiseksi. Oletetaan, että $\boldsymbol{\Theta}_0$ jokin parametriavaruuden $\boldsymbol{\Theta}$ osajoukko. Testataan hypoteesia

$$H_0: \ \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \quad \text{vs.} \quad H_1: \ \boldsymbol{\theta} \notin \boldsymbol{\Theta}_0.$$

Suurimman uskottavuuden testi hylkää H_0:n, jos

$$\Lambda = \frac{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta})} < c,$$

missä c on sopivasti valittu vakio.

5.7 Yksittäisten muuttujien testaaminen, kun H_0 on hylätty T^2–testillä

Jos hypoteesi $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ hylätään, niin $\boldsymbol{\mu}_{i1} \neq \boldsymbol{\mu}_{i2}$ ainakin yhdellä i:n arvolla $i = 1, 2, \ldots, p$. Mutta siitä huolimatta ei ole mitään takeita, että mikään hypoteeseista $H_0: \mu_{i1} = \mu_{i2}$ yksittäin testattuina tulisi hylätyksi. Tarkastellaan lineaarista yhdistettä $z = \mathbf{a}'\mathbf{y}$ ja testisuuretta

$$t(\mathbf{a}) = \frac{\bar{z}_1 - \bar{z}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_z^2}}. \tag{5.7.21}$$

Jos $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ hylätään, niin $H_0: \mathbf{a}'\boldsymbol{\mu}_1 = \mathbf{a}'\boldsymbol{\mu}_2$ hylätään ainakin yhdellä $\mathbf{a}$:n arvolla testisuureen (5.7.21) perusteella.

Lineaarisen yhdisteen ominaisuuksien nojalla $\bar{z} = \mathbf{a}'\bar{\mathbf{y}}$ ja $s_z^2 = \mathbf{a}'\mathbf{S}_{pl}\mathbf{a}$, missä $\mathbf{S}_{pl}$ on yhdistetty otoskovarianssimatriisi. Testisuure (5.7.21) voidaan siis kirjoittaa muodossa

$$t(\mathbf{a}) = \frac{\mathbf{a}'\bar{\mathbf{y}}_1 - \mathbf{a}'\bar{\mathbf{y}}_2}{\sqrt{\left(\frac{n_1+n_2}{n_1 n_2}\right) \mathbf{a}'\mathbf{S}_{pl}\mathbf{a}}}. \tag{5.7.22}$$

Jotta löydetään $\mathbf{a}$:n arvo, joka johtaa H_0:n hylkäämiseen, maksimoidaan $t^2(\mathbf{a})$ vektorin $\mathbf{a}$ suhteen. Näin löydetään $t(\mathbf{a})$:n maksimaalinen itseisarvo. Maksimi saavutetaan, kun

$$\mathbf{a} = \mathbf{S}_{pl}^{-1}(\mathbf{y}_1 - \mathbf{y}_2). \tag{5.7.23}$$

Kun $\mathbf{a} = \mathbf{S}_{pl}^{-1}(\mathbf{y}_1 - \mathbf{y}_2)$, niin funktiota $z = \mathbf{a}'\mathbf{y}$ kutsutaan erottelufunktioksi.

Jos T^2–testi hylkää hypoteesin $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, niin erottelufunktiota $\mathbf{a}'\mathbf{y}$ käyttäen hypoteesi $H_0: \mathbf{a}'\boldsymbol{\mu}_1 = \mathbf{a}'\boldsymbol{\mu}_2$ hylätään. Sen jälkeen voidaan tutkia yksittäisten kertoimien a_i suuruutta, joka kertoo y_i:n tärkeydestä H_0:n hylkäämisessä.

5.8 Kovarianssirakenteen testaus

5.8.1 Hypoteesi $H_0: \boldsymbol{\Sigma} = \sigma^2\mathbf{I}$

Kovarinssimatriisia $\boldsymbol{\Sigma}$ koskeva oletus $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ on tavanomainen esimerkiksi regresioanalyysin yhteydessä. Tarkastellaan siis nollahypoteesin $H_0: \boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ testaamista, kun vaihtoehtoinen hypoteesi on $H_1: \boldsymbol{\Sigma} \neq \sigma^2\mathbf{I}$. Olkoon $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ otos normaalijakaumasta $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Uskottavuussuhde hypoteesin $H_0: \boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ testaamiseksi on

$$\text{US} = \left[\frac{|\mathbf{S}|}{\text{tr}(\mathbf{S}/p)^p}\right]^{n/2}. \tag{5.8.24}$$

On osoitettu, että

$$-2\log(\mathrm{US}) \qquad \text{noudattaa likimain} \quad \chi^2(\nu), \tag{5.8.25}$$

kun n on suuri, missä ν on parametrien kokonaismäärä vähennettynä H_0:n vallitessa estimoitujen parametrien lukumäärällä ja log on luonnollinen logaritmi. Matriisissa $\mathbf{\Sigma}$ on $p(p+1)/2$ parametria ja estimoitaessa σ^2 menetetään yksi vapausaste, joten tässä $\nu = p(p+1)/2 - 1$.

5.8.2 Kovarianssimatriisien yhtäsuuruuden testaus

Nollahypoteesi k:n populaation kovarianssimatriisien yhtäsuuruuden testaamiseksi on muotoa

$$H_0: \ \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_k. \tag{5.8.26}$$

Erikoistapaus $H_0: \ \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ saadaan valitsemalla $k = 2$. Tehdään populaatioista riippumattomat otokset ja olkoot otoskoot $n_1, n_2, \ldots, n_k$. Testisuureen muodostamiseksi lasketaan

$$M = \frac{|\mathbf{S}_1|^{\nu_1/2}|\mathbf{S}_2|^{\nu_2/2}|\ldots|\mathbf{S}_k|^{\nu_k/2}}{|\mathbf{S}_{pl}|^{\nu/2}},$$

missä $\nu_i = n_i - 1$, $\nu = \sum_{i=1}^{k} \nu_i$, $\mathbf{S}_i$ on i. otoksesta laskettu otoskovarianssimatriisi ja

$$\mathbf{S}_{pl} = \sum_{i=1}^{k} \nu_i \mathbf{S}_i / \nu.$$

Silloin

$$-2(1-c)\log(M) \qquad \text{noudattaa likimain} \quad \chi^2[(k-1)p(p+1)/2], \tag{5.8.27}$$

missä

$$c = (\sum_{i=1}^{k} \frac{1}{\nu_i} - \frac{1}{\nu})[\frac{2p^2+3p-1}{6(p+1)(k-1)}].$$

5.8.3 Riippumattomuuden testaus

Oletetaan, että

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \sim N_{p+q}(\boldsymbol{\mu}, \mathbf{\Sigma}),$$

missä

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}.$$

Uskottavuussuhdetesti hypoteesille $H_0: \ \mathbf{\Sigma}_{12} = \mathbf{0}$ hylkää H_0:n vaihtoehtoa $H_1: \ \mathbf{\Sigma}_{12} \neq \mathbf{0}$ vastaan suurilla testisuureen

$$-2\log(US) = n\log(\frac{|\mathbf{S}_{11}||\mathbf{S}_{22}|}{|\mathbf{S}|})$$

arvoilla, missä

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix}.$$

Bartlett on osoittanut, että

$$-(n-1-(p+q+1)/2)\log(\frac{|\mathbf{S}_{11}||\mathbf{S}_{22}|}{|\mathbf{S}|}) \qquad \text{noudattaa likimain} \quad \chi^2(pq), \tag{5.8.28}$$

kun $n-(p+q)$ on suuri.