Datan ryhmittely

Harjoitustyö opintojaksolle DATA.STAT.450 Monimuuttujamenetelmät

Minttu Niiranen

02.03.2023

Sisällysluettelo

Johdanto

Tämä on Monimuuttujamenetelmät kurssin harjoitustyö. Harjoitustyössä tutkitaan valittua aineistoa graafisesti ja kahdella eri analyysin menetelmällä.

Harjoitustyön aineisto on datajoukko, jossa on mitattu kemiallisia ominaisuuksia useammasta eri viinistä. Datan on peräisin UCI arkistosta, joka löytyy lähdeluettelosta. Data on vuodelta 1991 ja se on luotu tarkoitukseen tunnistaa viinien alkuperä. Lähteestä löytyy myös datan tarkempia tietoja, kuten aineiston tekijä.

Dataa käsitellään harjoitustyössä kokonaisuudessaan, ja tarkoituksena on löytää datasta klustereita k-means menetelmällä eli tunnistaa datasta erilaisia ryhmiä niiden kemiallisten ominaisuuksien perusteella. Lisäksi tutkitaan aineistoa pääkomponenttianalyysillä.

Datan esittely

Data sisältää 178 havaintoa ja 14 muuttujaa, joista 13 viimeisintä kertoo viinien kemiallisista mitatuista ominaisuuksista. Tutkitaan aineistoa numeerisesti ja graafisesti.

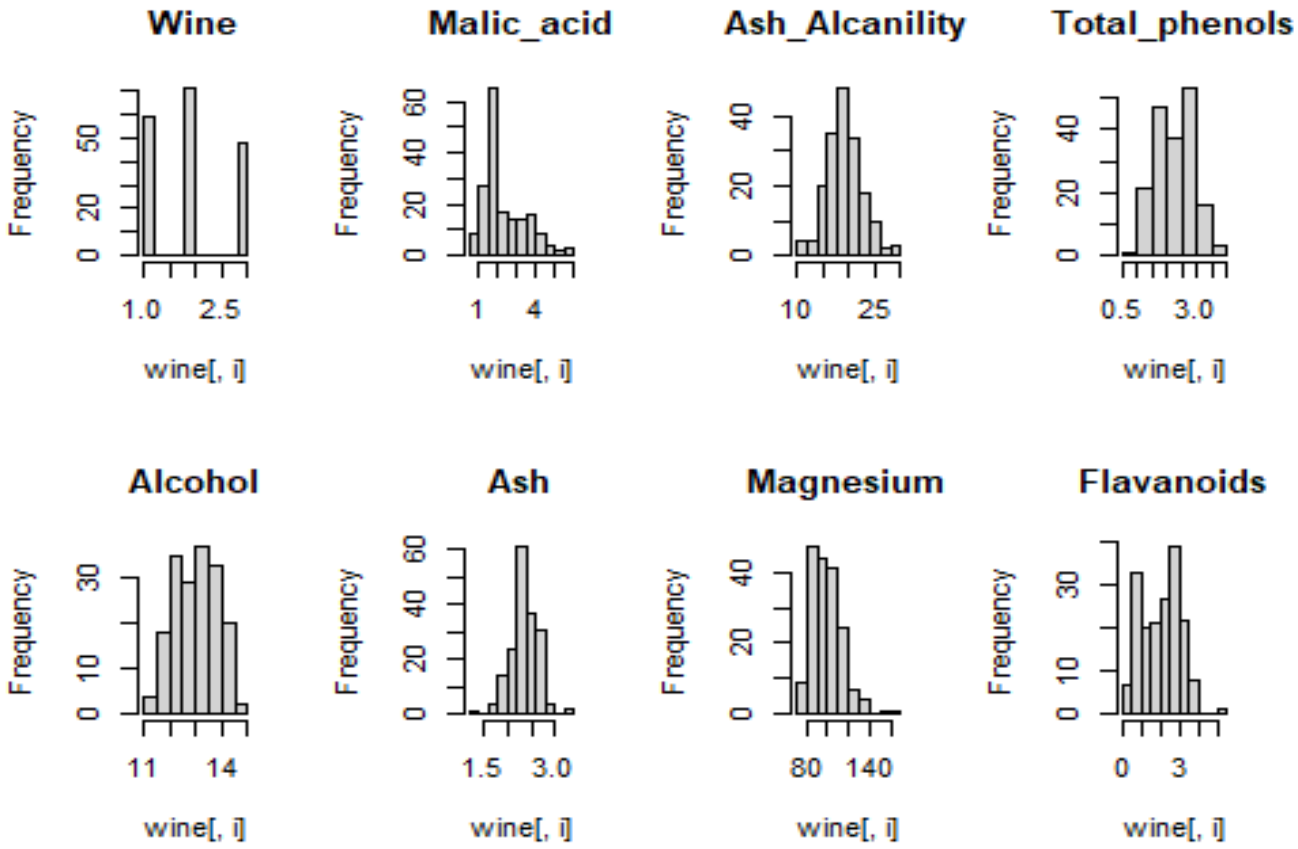
```
# Luetaan data tiedostosta
setwd("C:/Users/mintt/OneDrive/Tiedostot/Monimuuttujamenetelmät")
wine <- read.table("wine.data", header = TRUE, sep=",")

# Tarkistetaan aineiston koko
head(wine)
```

```
##   Wine Alcohol Malic_acid  Ash Ash_Alcanility Magnesium Total_phenols
## 1    1   14.23       1.71 2.43           15.6       127          2.80
## 2    1   13.20       1.78 2.14           11.2       100          2.65
## 3    1   13.16       2.36 2.67           18.6       101          2.80
## 4    1   14.37       1.95 2.50           16.8       113          3.85
## 5    1   13.24       2.59 2.87           21.0       118          2.80
## 6    1   14.20       1.76 2.45           15.2       112          3.27
##   Flavanoids Nonflavanoid_phenols Proanthocyanins Color_intensity  Hue   OD
## 1       3.06                0.28            2.29            5.64 1.04 3.92
## 2       2.76                0.26            1.28            4.38 1.05 3.40
## 3       3.24                0.30            2.81            5.68 1.03 3.17
## 4       3.49                0.24            2.18            7.80 0.86 3.45
## 5       2.69                0.39            1.82            4.32 1.04 2.93
## 6       3.39                0.34            1.97            6.75 1.05 2.85
##   Proline
## 1    1065
## 2    1050
## 3    1185
## 4    1480
## 5     735
## 6    1450

dim(wine)

## [1] 178  14

# Tutkitaan aineiston yhteenvetoa
summary(wine)

##       Wine          Alcohol        Malic_acid         Ash
##  Min.   :1.000   Min.   :11.03   Min.   :0.740   Min.   :1.360
##  1st Qu.:1.000   1st Qu.:12.36   1st Qu.:1.603   1st Qu.:2.210
##  Median :2.000   Median :13.05   Median :1.865   Median :2.360
##  Mean   :1.938   Mean   :13.00   Mean   :2.336   Mean   :2.367
##  3rd Qu.:3.000   3rd Qu.:13.68   3rd Qu.:3.083   3rd Qu.:2.558
##  Max.   :3.000   Max.   :14.83   Max.   :5.800   Max.   :3.230
##  Ash_Alcanility    Magnesium      Total_phenols     Flavanoids
##  Min.   :10.60   Min.   : 70.00   Min.   :0.980   Min.   :0.340
##  1st Qu.:17.20   1st Qu.: 88.00   1st Qu.:1.742   1st Qu.:1.205
##  Median :19.50   Median : 98.00   Median :2.355   Median :2.135
##  Mean   :19.49   Mean   : 99.74   Mean   :2.295   Mean   :2.029
##  3rd Qu.:21.50   3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.875
##  Max.   :30.00   Max.   :162.00   Max.   :3.880   Max.   :5.080
##  Nonflavanoid_phenols Proanthocyanins Color_intensity       Hue
##  Min.   :0.1300       Min.   :0.410   Min.   : 1.280   Min.   :0.4800
##  1st Qu.:0.2700       1st Qu.:1.250   1st Qu.: 3.220   1st Qu.:0.7825
##  Median :0.3400       Median :1.555   Median : 4.690   Median :0.9650
##  Mean   :0.3619       Mean   :1.591   Mean   : 5.058   Mean   :0.9574
##  3rd Qu.:0.4375       3rd Qu.:1.950   3rd Qu.: 6.200   3rd Qu.:1.1200
##  Max.   :0.6600       Max.   :3.580   Max.   :13.000   Max.   :1.7100
##        OD           Proline
```

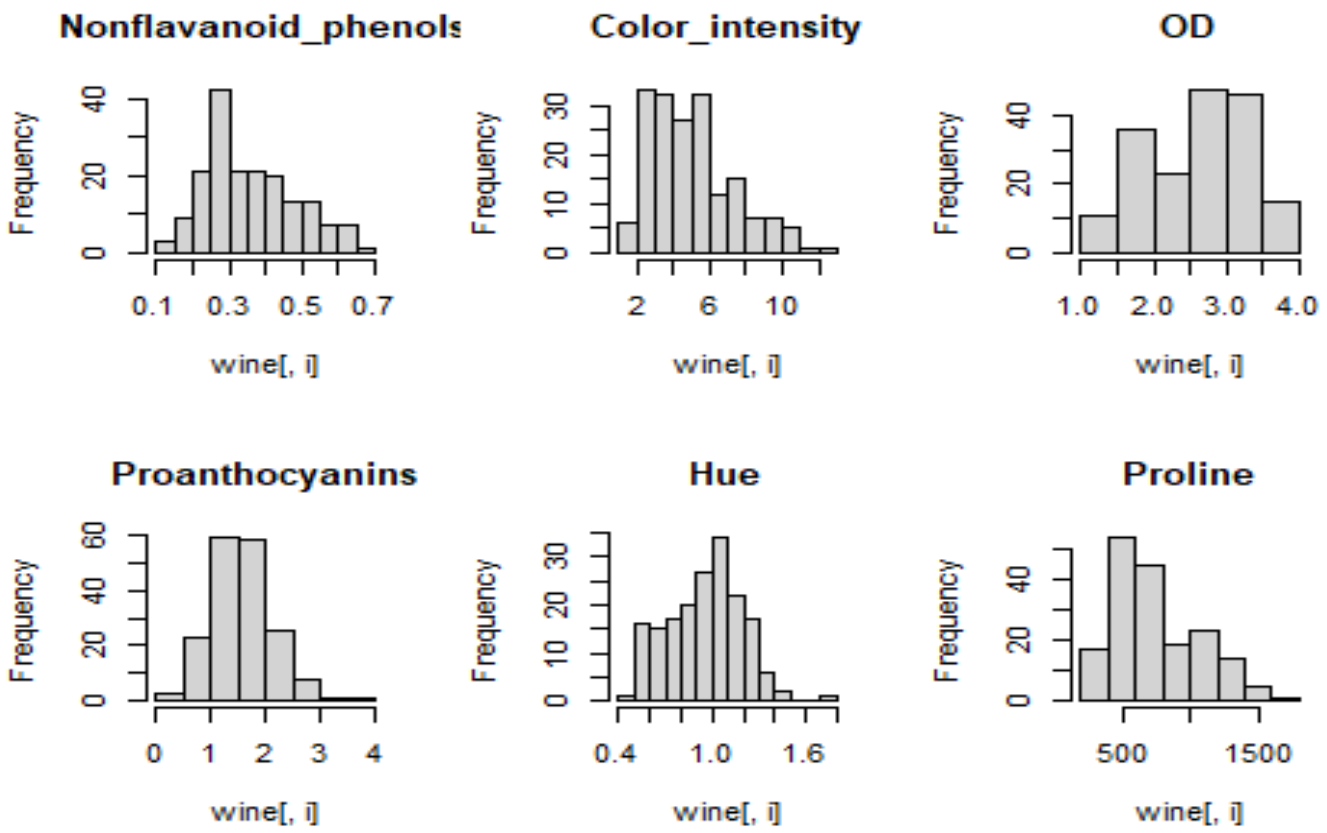
```
##  Min.   :1.270   Min.   : 278.0
##  1st Qu.:1.938   1st Qu.: 500.5
##  Median :2.780   Median : 673.5
##  Mean   :2.612   Mean   : 746.9
##  3rd Qu.:3.170   3rd Qu.: 985.0
##  Max.   :4.000   Max.   :1680.0
```

```
# Plotataan data
winetypes = wine[,1]
pairs(wine[,2:14], col=winetypes, lower.panel = NULL, cex.labels = 0.55)
```


```
# Plotataan histogrammit
layout(matrix(1:8, nc=4))
for(i in 1:8){hist(wine[,i], main=colnames(wine)[i])}
```




```
layout(matrix(1:6, nc=3))
for(i in 9:14){hist(wine[,i], main=colnames(wine)[i])}
```



Viinejä on tiivistelmän perusteella kolmea erilaista. Aineisto ei sisällä puuttuvia arvoja, joita pitäisi huomioida analyysivaiheessa. Histogrammeista näkee muuttujien esiintymismäärät datassa.

Plotatussa kuvassa on jokainen viini omalla värillään. Tarkastelemalla kuvia viinit menevät paikoitellen päällekkäin, joten klusterointi menetelmillä saadut klusterit voivat olla jaoteltu jollain muullakin tavalla ja niitä voi olla eri määrä. Tähän palataan myöhemmin Tulokset ja niiden tulkinta -kappaleessa.

Käytetyt menetelmät

Datan käsittely muuttuu haastavammaksi, kun havaintojen ja muuttujien määrä kasvaa suuremmaksi. Vastaavaan dataan ollaan kehitetty useita erilaisia menetelmiä, joilla datasta saadaan selville oleellisia tuloksia ja datan graafinen esittäminen helpottuu. Osa menetelmistä perustuu muun muassa datan dimensioiden vähentämiseen, jolloin sen käsitteleminen on yksinkertaisempaa.

Tässä harjoitustyössä tarkastellaan klusterointia k-means menetelmällä sekä pääkomponenttianalyysia, jossa pyritään edellä mainittuun dimensioiden vähentämiseen. Molemmista menetelmistä kerrotaan tarkemmin seuraavissa kappaleissa, mutta esitetään ensin yleinen teoria menetelmien taustalle.

Klusteroinnissa yleisesti pyritään jakamaan data ryhmiin eli klustereihin, joissa klusterin sisällä olevat havainnot ovat keskenään mahdollisimman samankaltaisia. Pyritään myös siihen, että klusterit ovat keskenään erilaisia. Klusterointi on yksi datan luokitteluun pyrkivä menetelmä ja siitä on useita erilasia versioita.

Pääkomponenttianalyysi pyrkii muodostamaan datasta uudet muuttujat eli pääkomponentit, jotka ovat alkuperäisten muuttujien lineaarisia kombinaatioita sekä korreloimattomia keskenään. Tarkoituksena on pääkomponenttien avulla selittää mahdollisimman paljon alkuperäisten muuttujien varianssista.

K-means ryhmittely

K-means klusterointi perustuu datan jakamiseen ennalta päätettyyn lukumäärään klustereita. Klusterin määrällä k jaetaan data ryhmiin $(G_1, G_2, \dots, G_k)$, missä G_i sisältää havainnot n_i, jossa $i = 1, .., k$ on klusterin järjestysluku.

Klusterien optimaalinen lukumäärä k saadaan minimoimalla jonkin numeerisen kriteerin suhteen, jossa pienet arvot viittavat parempaan ratkaisuun. Yleisimmin käytetty tapa on minimoida klusterin sisäinen neliösumma (within-group sum of squares, WGSS) kaikkien muuttujien yli.

Datan jakamiseen k-means menetelmällä löytyy omia algoritmeja, jotka yleensä alkavat jostakin alkuarvoisesta jaosta klustereihin, minkä jälkeen siirrytään iteratiivisesti askel kerrallaan kohti optimaalisempaa ratkaisua. Tällaiset algortimit eivät takaa globaalin optimaalisen ratkaisun löytämistä, vaan voivat myös päätyä lokaaliin ratkaisuun. Tässä harjoitustyössä käytetään R:n omia työkaluja, joten ei syvennytä algoritmien teoriaan tarkemmin.

Pääkomponenttianalyysi (PCA)

Pääkomponenttianalyysin tavoitteena on selittää korreloivien muuttujien$x^T = (x_1, \dots, x_q)$ varianssit uusien korreloimattomien muuttujien $y^T = (y_1, .., y_q)$ avulla, jotka ovat lineaarikombinaatioita alkueräisistä muuttujista.

Uudet muuttujat johdetaan laskevassa tärkeyden järjestyksessä siten, että y_1 selittää mahdollisimman paljon alkuperäisen datan varianssista kaikken mahdollisten lineaarikombinaatioiden joukosta. Sitten muodostetaan y_2 niin, että se selittää mahdollisimman paljon jäljellä olevasta varianssista ja on samalla korreloimaton $y_1 : n$ kanssa ja niin edelleen saadaan seuraavat pääkomponentti muuttujat. Se, miten komponentit löydetään, ei käsitellä tässä harjoitustyössä.

Tulokset ja niiden tulkinta

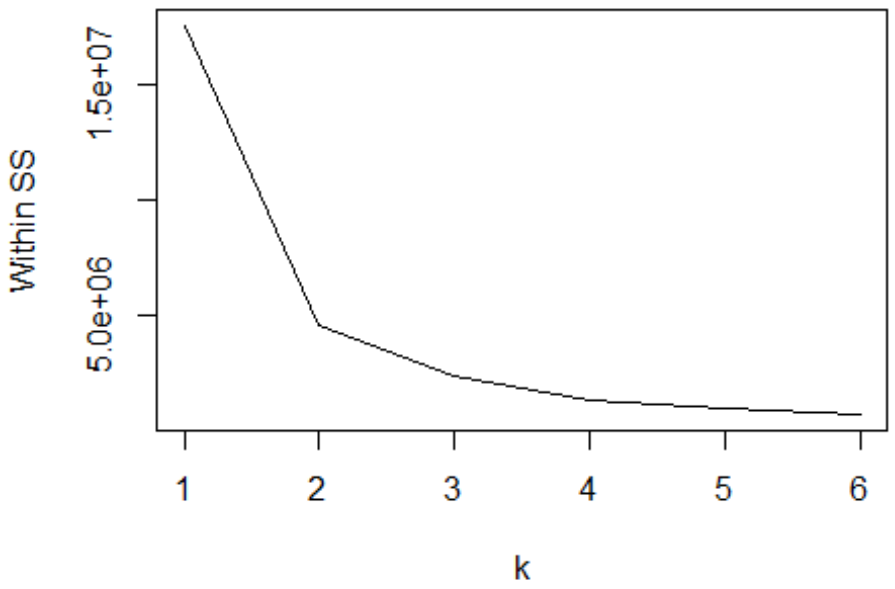
Tässä kappaleessa tutkitaan aineistoa valituilla menetelmillä ja tulkitaan tulokset ja menetelmien onnistuminen.

K-means ryhmittely

Tutkitaan aineistoa kmeans-menetelmällä. Klustereiden sopivaa määrää voidaan tutkia toistamalla kmeans menetelmä usealla eri klusterien määrällä. Klusteria kuvaava tunnusluku

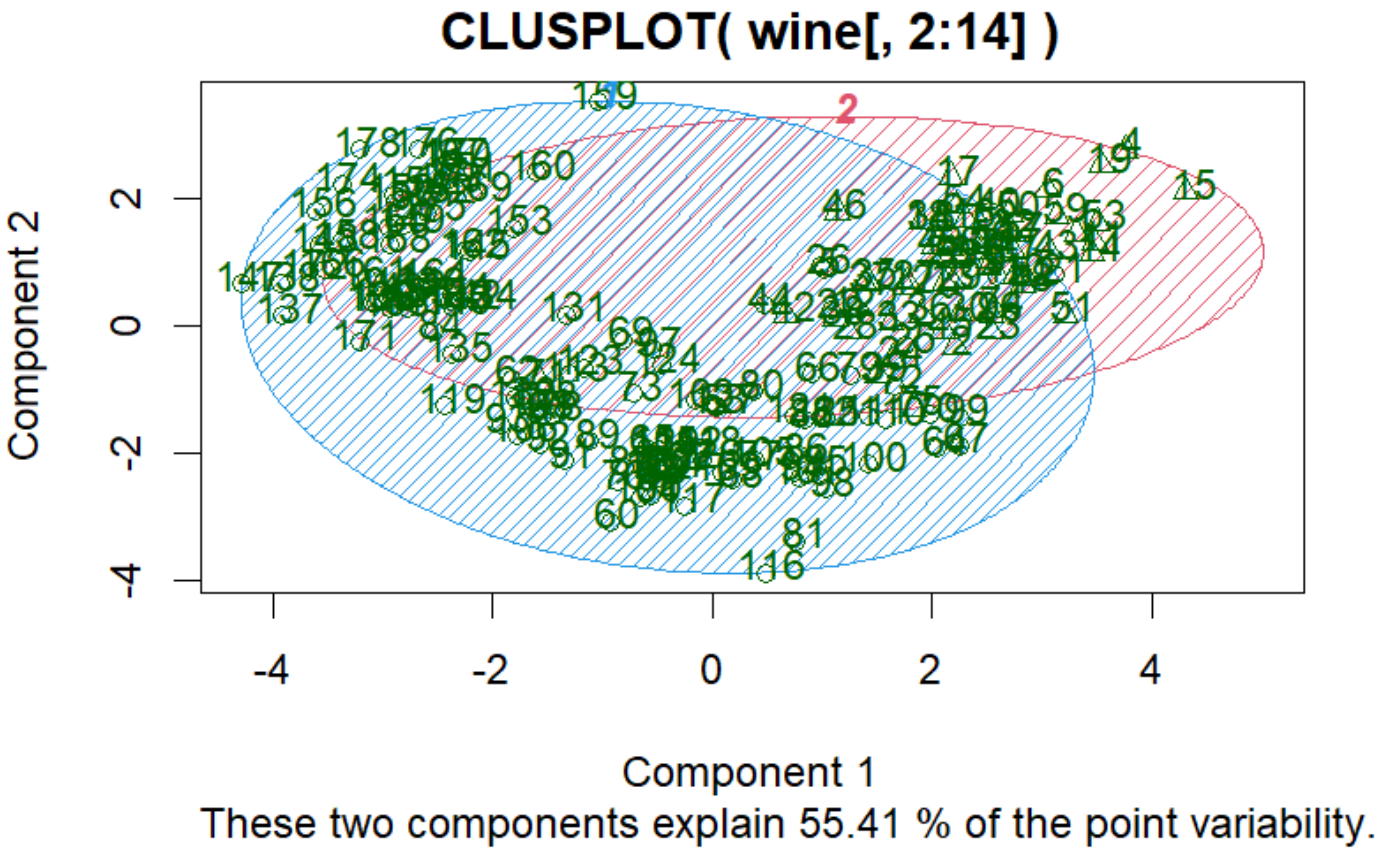
within_ss tarkoittaa klusterin sisäisiä neliösummia, jotka voidaan laskea yhteen. Kun klusteriin kuuluvat havainnot ovat lähellä toisiaan ja samanlaisia keskenään, klusterin sisällä oleva hajonta on pientä. Näin voidaan yhdellä tapaa päätellä sopiva määrä klustereita dataan.

```
ss <- rep(0,6); k <- 1:6
for (j in 1:6){ss[j] <- sum(kmeans(wine[,2:14],j)$withinss)}
plot(ss~k, ylab="Within SS", type="l")
```



Yllä olevasta kuvasta huomataan, että saavutetaan jo suuri etu kahdella klusterilla verrattuna yhteen klusteriin. Kasvattamalla klusterien määrää vielä suuremmaksi ei tapahdu suurta muutosta, joten valitaan tehdä klusterianalyysi kahdella klusterilla.

```
# Kmeans menetelmä
set.seed(10)
km<-kmeans(wine[,2:14],2)

# Piirretään kuva klustereista
library(cluster)
clusplot(wine[,2:14], km$cluster, color=TRUE, shade=TRUE,
labels=2, lines=0)
```




```
# Katsotaan, miten alkuperäiset viinit jakaantuvat klusterien kesken
table(km$cluster, winetypes)

##    winetypes
##      1  2  3
##   1  9 67 47
##   2 50  4  1
```

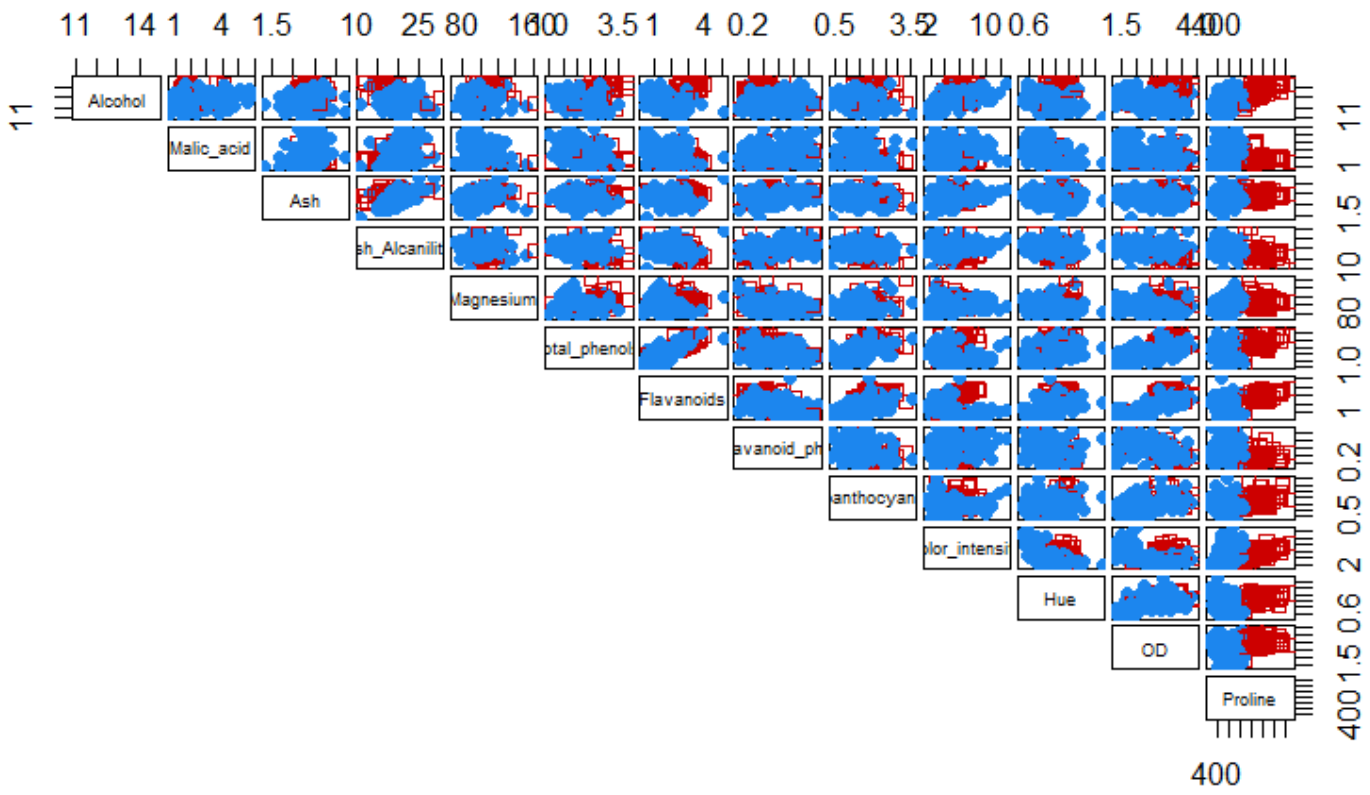
Table-komennolla saadun jaottelun mukaan viiniä 1 on eniten klusterissa 2 ja viinejä 2 ja 3 on eniten klusterissa 1. Jonkin verran havaintoja on jokaisesta viinistä mennyt myös toiseen klusteriin, joten jako ei ole yksiselitteinen. Voidaan kuitenkin päätellä, että viinit 2 ja 3 muistuttavat ominaisuuksiltaan enemmän toisiaan, kun taas viini 1 eroaa suuremmalta osin näistä kahdesta viinistä.

Clusplot-komennolla saadussa kuvaajassa klusterit menevät jonkin verran päällekkäin, kun data piirretään kaksiulotteisesti. Komponentit selittävät nyt noin 55% varianssista, mikä on yli puolet varianssista, ja loput datan varianssista jää selittämättä.

Kmeans tuloksista voidaan muodostaa uusi pisteparvi kuvaaja, kuten aineiston graafisessa esityksessä, mutta esitetään ne klusterien mukaan.

```
library(mclust)
```

```
## Package 'mclust' version 6.0.0
## Type 'citation("mclust")' for citing this R package in publications.

# Plotataan data
clPairs(wine[,2:14], km$cluster, cex.labels = 0.55, lower.panel = NULL)
```



Nähdään, että joissain kuvissa klusterit menevät päällekkäin. Suurin ero klusterien välillä nähdään proline-muuttujan (suom. proliini) sarakkeessa.

Tutkitaan vielä menetelmän tunnuslukuja tarkemmin.

```
print(km)

## K-means clustering with 2 clusters of sizes 123, 55
##
## Cluster means:
##    Alcohol Malic_acid      Ash Ash_Alcanility Magnesium Total_phenols
## 1 12.70285   2.544553 2.339106       20.40813  96.81301      2.062114
## 2 13.66655   1.870727 2.427818       17.45273 106.29091      2.816182
##   Flavanoids Nonflavanoid_phenols Proanthocyanins Color_intensity       Hue
## 1   1.641463            0.3926829        1.454065        4.851382 0.9086179
## 2   2.896545            0.2929091        1.896909        5.520364 1.0666545
##         OD   Proline
## 1 2.408211  565.8699
## 2 3.066727 1151.7273
```

```
##
## Clustering vector:
##   [1] 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 2 2 1 1 2 2 2 2 2 2 2 2 2 2
##  [38] 2 2 1 1 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 2 1 1 2
##  [75] 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [149] 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 2566028 1977722
##  (between_SS / total_SS =  74.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Tulostuksesta nähdään muun muassa klustereiden koot ja keskiarvot muuttujille klustereissa. Keskiarvojen ero on suurin proline-muuttujalle, muissa muuttujissa erot ovat suhteellisesti pienempiä, mutta kuitenkin selkeitä.

Voidaan tarkastella myös tunnuslukujen suhdetta between_SS/total_SS. Total_SS kuvaa neliösummaa koko datasta ja between_SS klustereiden välistä neliösummaa. Mitä suurempi osa koko datan hajonnasta voidaan selittää klusterien välisellä hajonnalla, sitä paremmin klusterit selittävät dataa.

Nyt arvo on noin 74%, mikä on reilusti yli puolet. Voidaan siis ajatella, että kaksi klusteria selittävät dataa kohtuullisen hyvin. Riippuu kuitenkin tilanteesta, halutaanko saavuttaa tätäkin parempi tulos ja missä hyvän tuloksen raja menee.

Pääkomponenttianalyysi

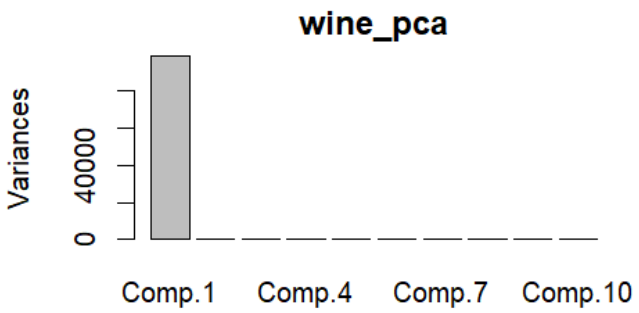
Tutkitaan seuraavaksi dataa pääkomponenttianalyysilla. Datassa on varsinaisia muuttujia 13, joten selvitetään voidaanko niistä saada järkeviä pääkomponentteja ja vähentää näin muuttujien määrää. Käytetään sekä kovarianssimatriisia ja korrelaatiomatriisia pääkomponenttianalyysissa ja esitetään niiden tulokset.

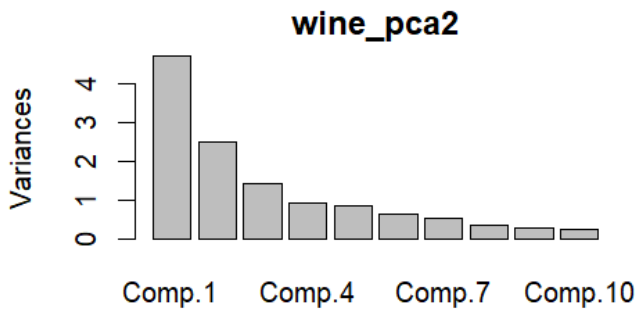
```
# Muodostetaan datasta sen kovarianssi ja korrelaatiomatriisi
S = cov(wine[,2:14])
R = cor(wine[,2:14])

# Suoritetaan pääkomponenttianalyysi
wine_pca <- princomp(covmat = S) #Kovarianssimatriisilla
wine_pca2 <- princomp(covmat = R) # Korrelaatiomatriisilla

# Tarkastellaan tiivistelmää
summary(wine_pca, loadings = TRUE)

## Importance of components:
##                           Comp.1       Comp.2       Comp.3       Comp.4
```

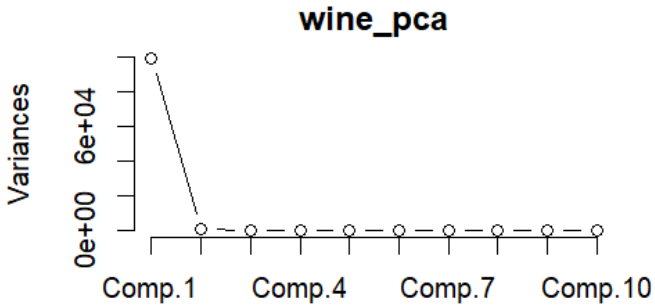
```
## Standard deviation     314.9631558 13.135268040 3.072151e+00 2.234095e+00
## Proportion of Variance   0.9980912  0.001735916 9.495896e-05 5.021736e-05
## Cumulative Proportion    0.9980912  0.999827146 9.999221e-01 9.999723e-01
##                             Comp.5       Comp.6       Comp.7       Comp.8
## Standard deviation     1.108533e+00 9.170953e-01 5.281794e-01 3.890775e-01
## Proportion of Variance 1.236368e-05 8.462130e-06 2.806815e-06 1.523081e-06
## Cumulative Proportion  9.999847e-01 9.999931e-01 9.999960e-01 9.999975e-01
##                             Comp.9      Comp.10      Comp.11      Comp.12
## Standard deviation     3.348085e-01 2.677734e-01 1.938452e-01 1.451632e-01
## Proportion of Variance 1.127830e-06 7.214158e-07 3.780603e-07 2.120138e-07
## Cumulative Proportion  9.999986e-01 9.999993e-01 9.999997e-01 9.999999e-01
##                            Comp.13
## Standard deviation     9.057430e-02
## Proportion of Variance 8.253928e-08
## Cumulative Proportion  1.000000e+00
##
## Loadings:
##                      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## Alcohol                                    0.141         0.194  0.923  0.285
## Malic_acid                          0.122  0.160 -0.613  0.742 -0.150
## Ash                                                                  -0.149
## Ash_Alcanility                      0.939 -0.331
## Magnesium                    0.999
## Total_phenols                                       0.315  0.279        -0.177
## Flavanoids                                    -0.169  0.525  0.434        -0.248
## Nonflavanoid_phenols
## Proanthocyanins                                     0.251  0.242 -0.310  0.870
## Color_intensity                     0.291  0.879  0.332        -0.113
## Hue
## OD                                            -0.178  0.261  0.289  0.102 -0.187
## Proline               1.000
##                      Comp.9 Comp.10 Comp.11 Comp.12 Comp.13
## Alcohol
## Malic_acid
## Ash                                -0.954   0.132  -0.174
## Ash_Alcanility
## Magnesium
## Total_phenols         0.256 -0.847
## Flavanoids            0.378  0.520   0.133
## Nonflavanoid_phenols                -0.199  -0.148   0.966
## Proanthocyanins                     -0.136
## Color_intensity
## Hue                                        -0.976  -0.167
## OD                   -0.874
## Proline

summary(wine_pca2, loadings = TRUE)

## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
```

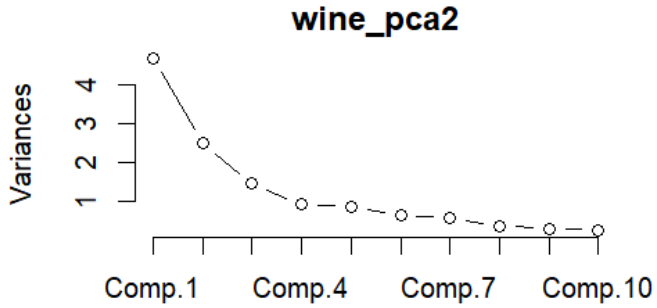
```
## Standard deviation     2.1692972 1.5801816 1.2025273 0.9586313 0.92370351
## Proportion of Variance 0.3619885 0.1920749 0.1112363 0.0706903 0.06563294
## Cumulative Proportion  0.3619885 0.5540634 0.6652997 0.7359900 0.80162293
##                            Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
## Standard deviation     0.80103498 0.74231281 0.59033665 0.53747553 0.50090167
## Proportion of Variance 0.04935823 0.04238679 0.02680749 0.02222153 0.01930019
## Cumulative Proportion  0.85098116 0.89336795 0.92017544 0.94239698 0.96169717
##                           Comp.11    Comp.12     Comp.13
## Standard deviation     0.47517222 0.41081655 0.321524394
## Proportion of Variance 0.01736836 0.01298233 0.007952149
## Cumulative Proportion  0.97906553 0.99204785 1.000000000
##
## Loadings:
##                       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## Alcohol                0.144  0.484  0.207         0.266  0.214         0.396
## Malic_acid            -0.245  0.225        -0.537         0.537 -0.421
## Ash                           0.316 -0.626  0.214  0.143  0.154  0.149 -0.170
## Ash_Alcanility        -0.239        -0.612                -0.101  0.287  0.428
## Magnesium              0.142  0.300 -0.131  0.352 -0.727        -0.323 -0.156
## Total_phenols          0.395        -0.146 -0.198  0.149                -0.406
## Flavanoids             0.423        -0.151 -0.152  0.109                -0.187
## Nonflavanoid_phenols -0.299        -0.170  0.203  0.501 -0.259 -0.595 -0.233
## Proanthocyanins        0.313        -0.149 -0.399 -0.137 -0.534 -0.372  0.368
## Color_intensity               0.530  0.137                -0.419  0.228
## Hue                    0.297 -0.279         0.428  0.174  0.106 -0.232  0.437
## OD                     0.376 -0.164 -0.166 -0.184  0.101  0.266
## Proline                0.287  0.365  0.127  0.232  0.158  0.120         0.120
##                       Comp.9 Comp.10 Comp.11 Comp.12 Comp.13
## Alcohol                0.509  0.212   0.226   0.266
## Malic_acid                   -0.309          -0.122
## Ash                   -0.308          0.499          -0.141
## Ash_Alcanility         0.200         -0.479
## Magnesium              0.271
## Total_phenols          0.286 -0.320  -0.304   0.304  -0.464
## Flavanoids                   -0.163                  0.832
## Nonflavanoid_phenols   0.196  0.216  -0.117           0.114
## Proanthocyanins       -0.209  0.134   0.237          -0.117
## Color_intensity              -0.291          -0.604
## Hue                          -0.522          -0.259
## OD                     0.137  0.524          -0.601  -0.157
## Proline               -0.576  0.162  -0.539
```

```
# Pääkomponenttien osuus
layout(matrix(1:4, nc = 2))
plot(wine_pca)
plot(wine_pca, type="line")
plot(wine_pca2)
plot(wine_pca2, type="line")
```









Pääkomponenttianalyyisssa voidaan tilaanteesta riippuen käyttää kovarianssimatriisia tai korrelaatiomatriisia. Pääkomponenttianalyysi ei ole skaalainvariantti, mikä on yksi sen haasteista. Lisäksi, jos alkuperäisten muuttujien varianssien välillä on suuria eroja ja yhden muuttujan varianssi on muita suurempi, voi muuttujalla olla alkuvaiheessa suuri vaikutus pääkomponentteihin. Kovarianssimatriisia on sopivin käyttää, kun muuttujat ovat suurin piirtein samaa skaalaa, ja muussa tapauksessa on hyvä käyttää korrelaatiomatriisia.

Kovarianssimatriisilla saadussa pääkomponenttianalyysin tuloksessa ensimmäinen komponentti selittää suuren osan varianssista osuudella 0.9980912 ja seuraavilla komponenteilla ei ole merkittävää vaikutusta.

Korrelaatiollamatriisilla saadussa tuloksessa neljän ensimmäisen komponentin kumulatiivinen osuus on 0.7359900. Usein valitaan komponenttien määrä siten, että ne selittävät noin 70-90% varianssista, joten tässä tapauksessa neljästä seitsemään pääkomponenttia.

Saatiin siis erilainen tulos verrattuna kovarianssimatriisilla saatuun tulokseen, jossa vastaukseksi saatiin yksi komponentti. Tulokset kovarianssi -ja korrelaatiomatriisilla eivät yleisesti ole toisiaan vastaavia ja siksi niillä tulee erilaisia tuloksia. Tässä tilanteessa korrelaatiomatriisilla saatu tulos antaa enemmän informaatiota ja yksittäinen muuttuja (esim. proline) ei pääse vaikuttamaan tulokseen liikaa.

Yhteenveto

Harjoitustyön tutkittavana aineistona oli kolmesta eri viinistä kerätyt havainnot niiden kemiallisista ominaisuuksista. Aluksi ryhmiteltiin aineistoa k-means menetelmällä ja saatiin vastaukseksi kaksi klusteria, jossa ensimmäisessä oli eniten viiniä 2 ja 3 ja toisessa klusterissa viiniä 1. Osa havainnoista meni kaikista viineistä myös toiseen klusteriin, joten jako ei ollut aivan tasainen. Tämä voi johtua monesta seikasta, aineistossa voi olla esimerkiksi sattumaa tai mittausvirhettä, jolloin jotkin havainnoista menivät väärään klusteriin. K-means menetelmä kuitenkin kohtalaisen hyvin onnistui datan ryhmittelyssä, kun tarkasteltiin tuloksia.

Pääkomponenttianalyysissa tavoitteena oli löytää pääkomponentit, joilla datan alkuperäiset 13 muuttujaa saataisiin esitettyä pienemmällä määrällä komponentteja. Pääkomponettianalyysi toteutettiin sekä kovarianssimatriisin ja korrelaatiomatriisin avulla, ja johtopäätöksenä todettiin, että aineiston perusteella korrelaatiomatriisilla saadaan pääkomponenttianalyysilla parempi tulos. Riippuen tilanteesta, data pystytään esittämään neljästä komponentista ylöspäin, joten onnistuttiin dimensioiden vähentämisessä.

Harjoitustyön menetelmät sopivat aineistoon kohtalaisen hyvin, mutta varsinkin k-means ryhmittelyssä tulos ei onnistunut täysin ja tuli virheellisiä luokitteluja. Virheet voivat johtua joko itse datasta tai ryhmittelyyn voidaan yrittää soveltaa jotain toista klusterointimenetelmää. K-means klusterointi on myös herkkä sellaisille havainnoille, jotka ovat irrallaan tai kaukana muusta datasta. Harjoitustyö keskittyy näihin kahteen menetelmään, joten muita mahdollisia menetelmiä datalle ei tutkittu.

Lähteet

https://archive.ics.uci.edu/ml/datasets/wine

Everitt, Brian., and Torsten. Hothorn. An Introduction to Applied Multivariate Analysis with R. 1st ed. 2011. New York, NY: Springer New York, 2011. Web.

Multivariate Analysis – luentomateriaali, Tapio Nummi

R dokumentaatio