PCA: Exploratory Data Analysis

Dr. H.G.J. van Mil (revision by Dr. F.J. Rodenburg)

September 2021

Save this file and "Day15.Rdata" in the same location. Then, after opening this file, go to **Session** > **Set Working Directory** > **To Source File Location**.

Then run this:

```
load("Day15.Rdata")
```

Use the presentation or the book to perform PCA on the data sets below (a complete example PCA analysis in R is shown in ***Elements of Biostatistics***, ***4.4***). Try to comment on the proportion of the variance captured by the first PCs, any clusters you might see, effects that seems to relate to certain groups, and try to improve the biplot with colors and/or symbols.

Protein sources in different nations

Description: These data measure protein consumption in twenty-five European countries for nine food groups. It is possible to use multivariate methods to determine whether there are groupings of countries and whether meat consumption is related to that of other foods.

Number of cases: $n = 25$
Variable names:

- `Country`: Country name
- `RdMeat`: Red meat
- `WhMeat`: White meat
- `Eggs`: Eggs
- `Milk`: Milk
- `Fish`: Fish
- `Cereal`: Cereal
- `Starch`: Starchy foods
- `Nuts`: Pulses, nuts, and oil-seeds
- `Fr&Veg`: Fruits and vegetables.

Can you interpret the results of this PCA analysis? Do certain countries cluster together and why?

```
summary(Protein)
```

```
##    Country             RedMeat         WhiteMeat          Eggs
##  Length:25          Min.   : 4.400   Min.   : 1.400   Min.   :0.500
##  Class :character   1st Qu.: 7.800   1st Qu.: 4.900   1st Qu.:2.700
##  Mode  :character   Median : 9.500   Median : 7.800   Median :2.900
```

```
##                     Mean   : 9.828   Mean   : 7.896   Mean   :2.936
##                     3rd Qu.:10.600   3rd Qu.:10.800   3rd Qu.:3.700
##                     Max.   :18.000   Max.   :14.000   Max.   :4.700
##       Milk            Fish           Cereals          Starch
##  Min.   : 4.90   Min.   : 0.200   Min.   :18.60   Min.   :0.600
##  1st Qu.:11.10   1st Qu.: 2.100   1st Qu.:24.30   1st Qu.:3.100
##  Median :17.60   Median : 3.400   Median :28.00   Median :4.700
##  Mean   :17.11   Mean   : 4.284   Mean   :32.25   Mean   :4.276
##  3rd Qu.:23.30   3rd Qu.: 5.800   3rd Qu.:40.10   3rd Qu.:5.700
##  Max.   :33.70   Max.   :14.200   Max.   :56.70   Max.   :6.500
##       Nuts           Fr.Veg
##  Min.   :0.700   Min.   :1.400
##  1st Qu.:1.500   1st Qu.:2.900
##  Median :2.400   Median :3.800
##  Mean   :3.072   Mean   :4.136
##  3rd Qu.:4.700   3rd Qu.:4.900
##  Max.   :7.800   Max.   :7.900
```

Morphological features of flea species

This data is from a paper by A. A. Lubischew, "On the Use of Discriminant Functions in Taxonomy", Biometrics, Dec 1962, pp.455–477.

The flea data set consist of 6 morphological features and the species:

- `tars1`, width of the first joint of the first tarsus in microns (the sum of measurements for both tarsi)
- `tars2`, the same for the second joint
- `head`, the maximal width of the head between the external edges of the eyes in 0.01 mm
- `ade1`, the maximal width of the aedeagus in the fore-part in microns
- `ade2`, the front angle of the aedeagus (1 unit = 7.5 degrees)
- `ade3`, the aedeagus width from the side in microns
- `species`, which species is being examined - concinna, heptapotamica, heikertingeri

Hypothesis about the data structure you find with PCA.

```
summary(Flea)
```

```
##      tars1           tars2            head           aede1
##  Min.   :122.0   Min.   :107.0   Min.   :43.00   Min.   :116.0
##  1st Qu.:148.0   1st Qu.:118.2   1st Qu.:49.00   1st Qu.:125.5
##  Median :185.5   Median :123.0   Median :50.50   Median :136.5
##  Mean   :177.3   Mean   :124.0   Mean   :50.35   Mean   :134.8
##  3rd Qu.:198.2   3rd Qu.:130.0   3rd Qu.:52.00   3rd Qu.:142.8
##  Max.   :242.0   Max.   :146.0   Max.   :58.00   Max.   :157.0
##      aede2           aede3          species
##  Min.   : 8.00   Min.   : 55.00   Length:74
##  1st Qu.:11.00   1st Qu.: 85.25   Class :character
##  Median :14.00   Median : 98.50   Mode  :character
##  Mean   :12.99   Mean   : 95.38
##  3rd Qu.:15.00   3rd Qu.:106.00
##  Max.   :16.00   Max.   :123.00
```

Fatty acid contents in olives from different regions

This data is from a paper by Forina, Armanino, Lanteri, Tiscornia (1983) Classification of Olive Oils from their Fatty Acid Composition, in Martens and Russwurm (ed) Food Research and Data Anlysis.

- `region` Three super-classes of Italy: North, South and the island of Sardinia
- `area` Nine collection areas: three from North, four from South and 2 from Sardinia
- `palmitic`, `palmitoleic`, `stearic`, `oleic`, `linoleic`, `linolenic`, `arachidic`, `eicosenoic` fatty acids percent ($\times 100$)

Form a hypothesis about the data structure you find with PCA, that could be studied with further research.

```
summary(Olive)
```

```
##      Region          Area              palmitic     palmitoleic
##  Min.   :1.000   Length:572         Min.   : 610   Min.   : 15.00
##  1st Qu.:1.000   Class :character   1st Qu.:1095   1st Qu.: 87.75
##  Median :1.000   Mode  :character   Median :1201   Median :110.00
##  Mean   :1.699                      Mean   :1232   Mean   :126.09
##  3rd Qu.:3.000                      3rd Qu.:1360   3rd Qu.:169.25
##  Max.   :3.000                      Max.   :1753   Max.   :280.00
##     stearic          oleic         linoleic        linolenic
##  Min.   :152.0   Min.   :6300   Min.   : 448.0   Min.   : 0.00
##  1st Qu.:205.0   1st Qu.:7000   1st Qu.: 770.8   1st Qu.:26.00
##  Median :223.0   Median :7302   Median :1030.0   Median :33.00
##  Mean   :228.9   Mean   :7312   Mean   : 980.5   Mean   :31.89
##  3rd Qu.:249.0   3rd Qu.:7680   3rd Qu.:1180.8   3rd Qu.:40.25
##  Max.   :375.0   Max.   :8410   Max.   :1470.0   Max.   :74.00
##    arachidic       eicosenoic
##  Min.   :  0.0   Min.   : 1.00
##  1st Qu.: 50.0   1st Qu.: 2.00
##  Median : 61.0   Median :17.00
##  Mean   : 58.1   Mean   :16.28
##  3rd Qu.: 70.0   3rd Qu.:28.00
##  Max.   :105.0   Max.   :58.00
```