Multivariate Statistical Analysis - Exercise Session 8

10.03.2022

Problem 1: Canonical correlation analysis

First we read the data and define groups X and Y.

```
car <- read.table("CAR.txt", header = TRUE, sep = "\t")
dim(car)
```

```
## [1] 24 10
```

```
head(car)
```

```
##      Type    Model Economy Service Value Price Design Sport Safety Easy.h.
## 1    Audi      100     3.9     2.8   2.2   4.2    3.0   3.1    2.4     2.8
## 2     BMW 5 series     4.8     1.6   1.9   5.0    2.0   2.5    1.6     2.8
## 3 Citroen       AX     3.0     3.8   3.8   2.7    4.0   4.4    4.0     2.6
## 4 Ferrari              5.3     2.9   2.2   5.9    1.7   1.1    3.3     4.3
## 5    Fiat      Uno     2.1     3.9   4.0   2.6    4.5   4.4    4.4     2.2
## 6    Ford   Fiesta     2.3     3.1   3.4   2.6    3.2   3.3    3.6     2.8
```

```
# X = (Price, Value)
x <- as.matrix(car[, c(6, 5)])

# Y = (Economy, Service, Desing, Sport, Safety, Easy.h)
y <- as.matrix(car[, c(3, 4, 7:10)])

xy <- cbind(x, y)
rownames(xy) <- paste(car$Type, car$Model)
head(xy)
```

```
##              Price Value Economy Service Design Sport Safety Easy.h.
## Audi 100       4.2   2.2     3.9     2.8    3.0   3.1    2.4     2.8
## BMW 5 series   5.0   1.9     4.8     1.6    2.0   2.5    1.6     2.8
## Citroen AX     2.7   3.8     3.0     3.8    4.0   4.4    4.0     2.6
## Ferrari        5.9   2.2     5.3     2.9    1.7   1.1    3.3     4.3
## Fiat Uno       2.6   4.0     2.1     3.9    4.5   4.4    4.4     2.2
## Ford Fiesta    2.6   3.4     2.3     3.1    3.2   3.3    3.6     2.8
```

a) Compute the sample canonical vectors with the corrected scaling

Let

$$z = (x^T, y^T)^T,$$

and let

$$\mathrm{Cov}(z) = \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Define

$$M_1 = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \quad \text{and}$$
$$M_2 = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

Now the canonical vectors α_k are the eigenvectors of M_1 (α_k corresponds to the kth largest eigenvalue), and the canonical vectors β_k are the eigenvectors of M_2. First we compute M_1 and M_2.

```
r <- cov(xy)
r11 <- r[1:2, 1:2] # cov(X)
r22 <- r[3:8, 3:8] # cov(Y)
r21 <- r[3:8, 1:2] # cov(Y,X)
r12 <- r[1:2, 3:8] # cov(X,Y)
r11_inv <- solve(r11)
r22_inv <- solve(r22)

m1 <- r11_inv %*% r12 %*% r22_inv %*% r21
m2 <- r22_inv %*% r21 %*% r11_inv %*% r12
```

Now we can compute *unscaled* canonical vectors.

```
alpha1 <- eigen(m1)$vectors[, 1]
alpha2 <- eigen(m1)$vectors[, 2]

beta1 <- eigen(m2)$vectors[, 1]
beta2 <- eigen(m2)$vectors[, 2]
```

We want to scale canonical vectors such that

$$\mathrm{Var}(\alpha_k^T x) = 1 = \mathrm{Var}(\beta_k^T y).$$

- How to get the correct scales?

Let $\tilde{\alpha}_k$ denote unscaled canonical vector. Then we get correctly scaled canonical vector α_k by

$$\alpha_k = \frac{1}{\mathrm{Std}(\tilde{\alpha}_k^T x)} \tilde{\alpha}_k,$$

since now

$$\mathrm{Var}(\alpha_k x) = \mathrm{Var}\left(\frac{1}{\mathrm{Std}(\tilde{\alpha}_k^T x)} \tilde{\alpha}_k x\right) = \left(\frac{1}{\mathrm{Std}(\tilde{\alpha}_k^T x)}\right)^2 \mathrm{Var}(\tilde{\alpha}_k^T x) = 1.$$

Additionally, by affine equivariance we have

$$\mathrm{Var}(\tilde{\alpha}_k^T x) = \tilde{\alpha}_k^T \mathrm{Cov}(x) \tilde{\alpha}_k = \tilde{\alpha}_k^T \Sigma_{11} \tilde{\alpha}_k$$
$$\Rightarrow \mathrm{Std}(\tilde{\alpha}_k^T x) = \sqrt{\tilde{\alpha}_k^T \Sigma_{11} \tilde{\alpha}_k}.$$

Similar calculations can be performed for β_k. Thus correctly scaled canonical vectors are given by

```
alpha1 <- alpha1 / sqrt((alpha1 %*% r11 %*% alpha1)[1, 1])
alpha2 <- alpha2 / sqrt((alpha2 %*% r11 %*% alpha2)[1, 1])
beta1 <- beta1 / sqrt((beta1 %*% r22 %*% beta1)[1, 1])
beta2 <- beta2 / sqrt((beta2 %*% r22 %*% beta2)[1, 1])
```

b) Score vectors

Sample scores can be calculated as

$$\eta_{ki} = \alpha_k^T x_i \quad \text{and } \phi_{ki} = \beta_k^T y_i,$$

where x_i is the ith row of X and y_i is the ith row of Y. More coincisely,

$$\eta_k = X\alpha_k \quad \text{and } \phi_k = Y\beta_k.$$

```
eta1 <- x %*% alpha1
eta2 <- x %*% alpha2
phi1 <- y %*% beta1
phi2 <- y %*% beta2
```

Now we can check that

$$\mathrm{Var}(\eta_k) = 1 = \mathrm{Var}(\phi_k),$$

$$\begin{cases} \mathrm{Cor}(\eta_1, \eta_2) = 0, \\ \mathrm{Cor}(\phi_1, \phi_2) = 0 \end{cases}$$

and that

$$\begin{cases} \mathrm{Cor}(\eta_1, \phi_1) = \sqrt{\lambda_1}, \\ \mathrm{Cor}(\eta_2, \phi_2) = \sqrt{\lambda_2}. \end{cases}$$

```
c(var(eta1), var(eta2), var(phi1), var(phi2))
```

```
## [1] 1 1 1 1
```

```
c(cor(eta1, eta2), cor(phi1, phi2))
```

```
## [1] -1.018901e-14 -5.058133e-16
```

```
c(cor(eta1, phi1), cor(eta2, phi2))
```

```
## [1] 0.9793946 0.9056556
```

```
sqrt(eigen(m1)$values)
```

```
## [1] 0.9793946 0.9056556
```

```
sqrt(eigen(m2)$values[1:2])
```

```
## [1] 0.9793946 0.9056556
```

So canonical correlations are $\rho_1 = 0.98$ and $\rho_1 = 0.91$.

c) Interpret the first pair of canonical variables

For the first pair of canonical variables we got

$$\eta_1 = 0.32 \times \textbf{Price} - 0.62 \times \textbf{Value}$$
$$\phi_1 = 0.43 \times \textbf{Economy} - 0.21 \times \textbf{Service} + 0 \times \textbf{Design} - 0.47 \times \textbf{Sport} - 0.22 \times \textbf{Safety} - 0.4 \times \textbf{Easy h.}.$$

Remember that scales for the variables are

$$1 = \text{very good and } 6 = \text{very bad}.$$

For example, **Value** $= 1$ means that the car loses its value slowly, which is a good thing. On the contrary, cars with **Value** $= 6$ lose value very fast.

First let's interpret x-axis of Figure 1. Based on the weights for **Price** and **Value** we have very expensive but valuable cars on the right. On the other hand, cheap cars that lose value fast are on the left. Note also, that **Value** has almost twice as much weight in the scores compared to **Price**. All in all, x-axis could be interpreted as *Value index of the car* and cars on the right can be considered as worthy investments.

We can interpret y-axis similarly. Variable **Design** has almost negligible weight. However, **Economy** has positive contribution to scores and other variables have negative weights. Thus uppermost cars on Figure 1 have very good **Service**, **Sport**, **Safety** and **Easy h.** but bad **Economy**. Therefore, uppermost cars use a lot of fuel but have otherwise good qualities (vice verce for lowest cars). All in all, y-axis can be interpreted as *Quality of the car*.

```
plot(eta1, phi1, xlab = expression(eta[1]),
       ylab = expression(phi[1]), pch = NA)
text(eta1, phi1, labels = paste(car$Type, car$Model))
```

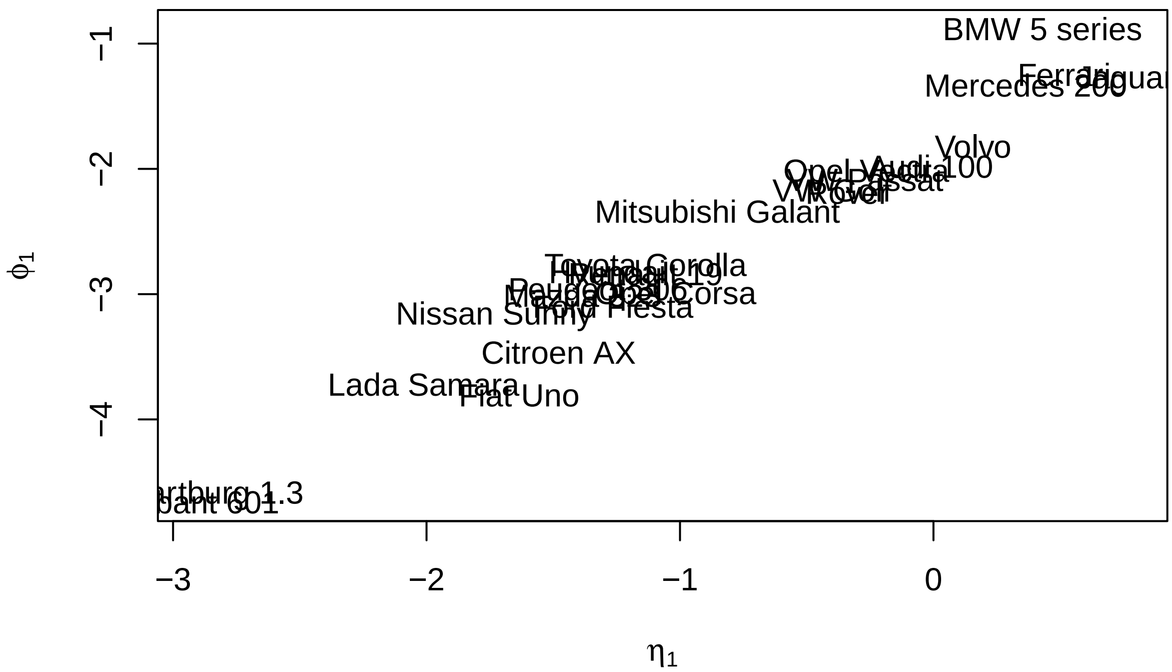

Figure 1: Scores corresponding to the first pair of canonical variables.

d) Interpret the second pair of canonical variables

For the second pair of canonical variables we got

$\eta_2 = -1.41 \times \textbf{Price} - 1.42 \times \textbf{Value}$

$\phi_2 = -0.46 \times \textbf{Economy} - 0.7 \times \textbf{Service} + 0.06 \times \textbf{Design} + 0 \times \textbf{Sport} + 0.3 \times \textbf{Safety} - 1.01 \times \textbf{Easy h.}.$

Now let's interpret x-axis of Figure 2. Notice that **Price** and **Value** have almost identical weights. Thus x-axis describes mean of Price and **Value**. So one interpretation for x-axis could be a simple *value for money index*. Cars one the right have very good value for money index but cars on the left have a poor value for money index.

```
# The x-axis reflects the average of Value and Price
sort(rowSums(xy[, 1:2]) / 2)
```

```
##          VW Golf        VW Passat          Hyundai       Opel Corsa
##             2.30             2.65             2.70             2.80
```

```
##       Opel Vectra        Ford Fiesta       Nissan Sunny             Volvo 
##              2.95               3.00               3.00               3.05 
##       Lada Samara       Mercedes 200          Audi 100        Peugeot 306 
##              3.15               3.20               3.20               3.20 
##        Renault 19     Toyota Corolla         Citroen AX          Mazda 323 
##              3.20               3.20               3.25               3.25 
##          Fiat Uno             Rover  Mitsubishi Galant       BMW 5 series 
##              3.30               3.30               3.35               3.45 
##       Trabant 601            Jaguar       Wartburg 1.3            Ferrari 
##              3.50               3.55               3.60               4.05 
```

Now the scores ϕ_2 reflect what kind of qualities cars with certain value for money index have. Weights for **Economy**, **Service** and **Easy h.** are negative, and weight for **Safety** is positive. On the other hand, weights for **Design** and **Sport** are negligible or very close to zero. Thus cars with good value for money index have good services, use little gas, are easy to handle but have maybe a bit worse safety than high-end cars. Below we show the raw numbers for an expensive car, a car with good value for money index and a cheap car.

```r
car[c(4, 22, 24), ]
```

```
##         Type Model Economy Service Value Price Design Sport Safety Easy.h.
## 4    Ferrari           5.3     2.9   2.2   5.9    1.7   1.1    3.3     4.3
## 22        VW  Golf     2.4     2.1   2.0   2.6    3.2   3.1    3.1     1.6
## 24 Wartburg   1.3      3.7     4.7   5.5   1.7    4.8   5.2    5.5     4.0
```

All in all, one interpretation for y-axis could be *consumer-friendliness*.

```r
plot(eta2, phi2, xlab = expression(paste(eta[2])),
       ylab = expression(paste(phi[2])), pch = NA)
text(eta2, phi2, labels = paste(car$Type, car$Model))
```

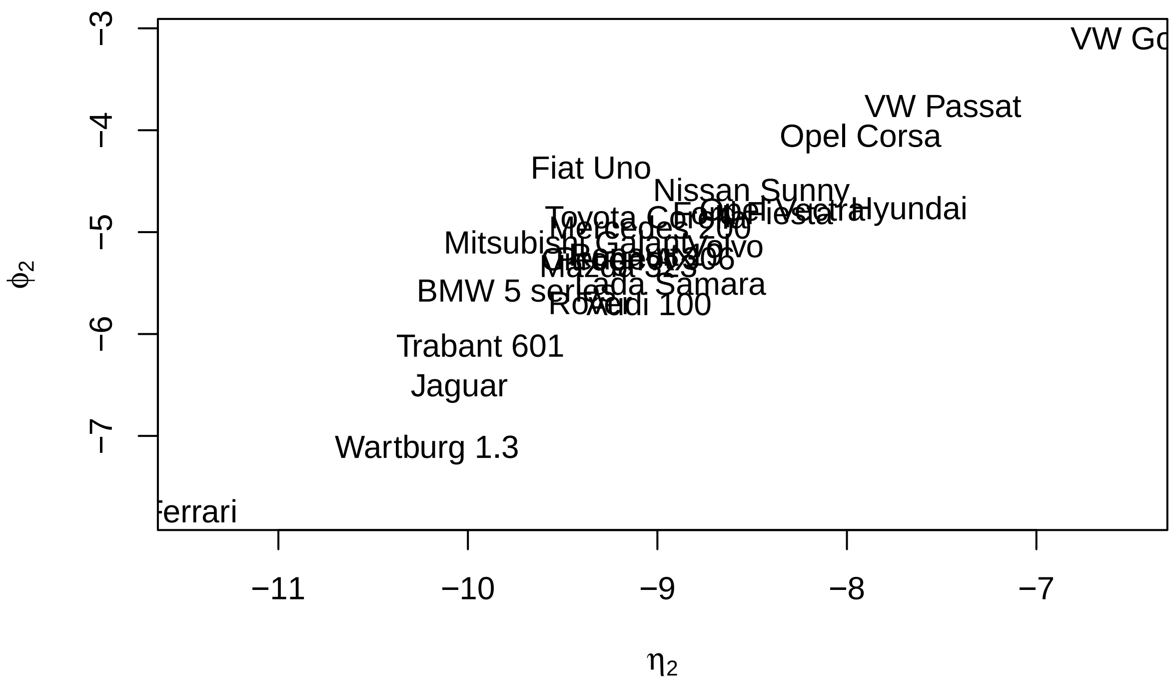

Figure 2: Scores corresponding to the second pair of canonical variables.

Problem 2: Discussion about course project

- Project work is compulsory and late submissions are not graded.
- See paragraph "About grading of the project work" in MyCourses Assignments tab. There you can find how the project is graded.
  - For example, If your project work is amazing in other parts but you do not include univariate analysis ⇒ 5/6p.
- You can use multiple methods that are learned on the course, however, it is suggested to use only one method for multivariate analysis.
- You can include code in the final pdf but it is not necessary.
- **No finding is a finding!**
- Some possible data sources are given below:
  - Kaggle
  - OECD
  - Statistics Finland
  - Our World in Data

Hint for homework 8

Ghost imaginary parts from eigenvectors can be removed with the function `as.numeric`.

```r
v1 <- c(1 + 0i, 2 + 0i, 3 + 0i)
class(v1)
```

```
## [1] "complex"
```

```r
v1
```

```
## [1] 1+0i 2+0i 3+0i
```

```r
v2 <- as.numeric(v1)
class(v2)
```

```
## [1] "numeric"
```

```r
v2
```

```
## [1] 1 2 3
```