Principale-Componenten Analyse

Oefening1: De onderstaande gegevens met betrekking tot de gemiddelde proteïne consumptie komende van verschillende voedselbronnen door inwoners in 25 Europese landen, kan je terug vinden in het bestand “Protein consumption in 25 European countries.csv”.

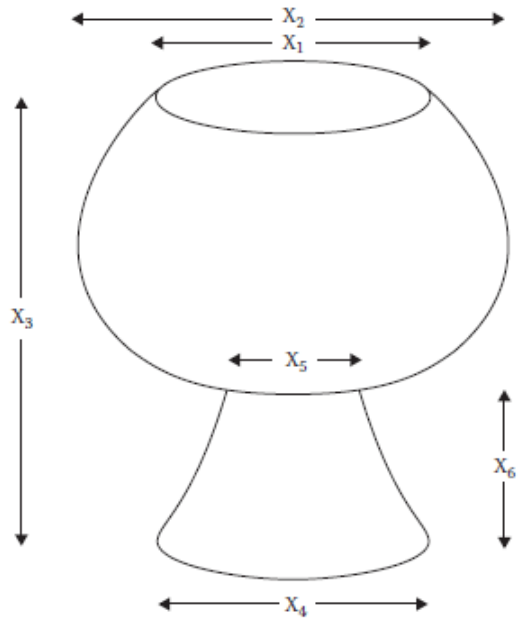
Table 6.7 Protein consumption (grams per person per day) in 25 European countries

Country	Red meat	White meat	Eggs	Milk	Fish	Cereals	Starchy foods	Pulses, nuts, and oilseeds	Fruit and vegetables	Total
Albania	10	1	1	9	0.0	42	1	6	2	72
Austria	9	14	4	20	2.0	28	4	1	4	86
Belgium	14	9	4	18	5.0	27	6	2	4	89
Bulgaria	8	6	2	8	1.0	57	1	4	4	91
Czechoslovakia	10	11	3	13	2.0	34	5	1	4	83
Denmark	11	11	4	25	10.0	22	5	1	2	91
East Germany	8	12	4	11	5.0	25	7	1	4	77
Finland	10	5	3	34	6.0	26	5	1	1	91
France	18	10	3	20	6.0	28	5	2	7	99
Greece	10	3	3	18	6.0	42	2	8	7	99
Hungary	5	12	3	10	0.0	40	4	5	4	83
Ireland	14	10	5	26	2.0	24	6	2	3	92
Italy	9	5	3	14	3.0	37	2	4	7	84
The Netherlands	10	14	4	23	3.0	22	4	2	4	86
Norway	9	5	3	23	10.0	23	5	2	3	83
Poland	7	10	3	19	3.0	36	6	2	7	93
Portugal	6	4	1	5	14.0	27	6	5	8	76
Romania	6	6	2	11	1.0	50	3	5	3	87
Spain	7	3	3	9	7.0	29	6	6	7	77
Sweden	10	8	4	25	8.0	20	4	1	2	82
Switzerland	13	10	3	24	2.0	26	3	2	5	88
United Kingdom	17	6	5	21	4.0	24	5	3	3	88
USSR	9	5	2	17	3.0	44	6	3	3	92
West Germany	11	13	4	19	3.0	19	5	2	4	80
Yugoslavia	4	5	1	10	1.0	56	3	6	3	89

Bron: Manly, Bryan and Navarro Alberto , Jorge (2017): Multivariate Statistical Methods A Primer, Fourth Edition, CRC Press

a. Plaats de gegevens in een dataframe met de naam “protein”.
b. Geef de rijen de naam van de landen die in de eerste kolom (variabele) staan (zet de namen van de landen dus als index)
c. Welke kolommen dien je te verwijderen om een Principale-Componenten Analyse te kunnen uitvoeren? Verwijder eventuele kolommen met de verkeerde meetschaal, kolommen die niet zinvol zijn,….
d. Bekijk de correlaties tussen de variabelen. Zijn de gegevens bruikbaar om er een Principale-Componenten Analyse op toe te passen?
e. Voer een Principale-Componenten Analyse uit en interpreteer de resultaten. Hoeveel componenten moet je gebruiken om 90% van de variantie te kunnen verklaren?
f. Zijn er een of meerdere variabelen die uitgesproken doorwegen in het bepalen van de eerste hoofdcomponent? Zo ja welke?
g. Maak een biplot
h. Maak een nieuw dataframe aan waarbij je voor de observaties de eerste drie hoofdcomponenten nemen.
i. Pas hier een hiërarchische cluster analyse op toe (euclidische afstand). Maak 6 clusters en kijk welke landen in elke cluster zitten. Vergelijk de resultaten met de resultaten bekomen met een cluster analyse toegepast op de oorspronkelijke variabelen.

Oefening 2: De onderstaande gegevens met betrekking tot prehistorische bekers (aardewerk) uit Thailand, kan je terug vinden in het bestand “Prehistoric goblets from Thailand.csv”.



Goblet	X_1	X_2	X_3	X_4	X_5	X_6
1	13	21	23	14	7	8
2	14	14	24	19	5	9
3	19	23	24	20	6	12
4	17	18	16	16	11	8
5	19	20	16	16	10	7
6	12	20	24	17	6	9
7	12	19	22	16	6	10
8	12	22	25	15	7	7
9	11	15	17	11	6	5
10	11	13	14	11	7	4
11	12	20	25	18	5	12
12	13	21	23	15	9	8
13	12	15	19	12	5	6
14	13	22	26	17	7	10
15	14	22	26	15	7	9
16	14	19	20	17	5	10
17	15	16	15	15	9	7
18	19	21	20	16	9	10
19	12	20	26	16	7	10
20	17	20	27	18	6	14
21	13	20	27	17	6	9
22	9	9	10	7	4	3
23	8	8	7	5	2	2
24	9	9	8	4	2	2
25	12	19	27	18	5	12

Bron: Manly, Bryan and Navarro Alberto , Jorge (2017): Multivariate Statistical Methods A Primer, Fourth Edition, CRC Press

a. Plaats de gegevens in het dataframe *goblet*
b. Welke kolommen dien je te verwijderen om een Principale-Componenten Analyse te kunnen uitvoeren? Verwijder eventuele kolommen met de verkeerde meetschaal, kolommen die niet zinvol zijn,….
c. Voer een Principale-Componenten Analyse uit en interpreteer de resultaten. Hoeveel componenten moet je gebruiken om 95% van de variantie te kunnen verklaren?
d. Maak een biplot
e. Maak een nieuw dataframe *verhoudingen* aan met de volgende variabelen:
   V1 = X2/X1
   V2 = X2/X4
   V3 = X1/X4
   V4 = X2/X5
   V5 = X1/X5
   V6 = X4/X5
   V7 = X3/X6
f. Voer een Principale-Componenten Analyse uit op het dataframe *verhoudingen* en interpreteer de resultaten. Hoeveel componenten moet je nu gebruiken om 95% van de variantie te kunnen verklaren?
g. Maak een biplot
h. Vergelijk de resultaten van beide Principale Componenten Analyse en formuleer een conclusie. Kijk daarbij naar de gemiddelde correlaties.

Oefening 3: Zoek de dataset “cpus.csv” en plaats die in een dataframe.

Ter info:

`cpus` *Performance of Computer CPUs*

Description

A relative performance measure and characteristics of 209 CPUs.

Usage

```
cpus
```

Format

The components are:

`name` manufacturer and model.

`syct` cycle time in nanoseconds.

`mmin` minimum main memory in kilobytes.

`mmax` maximum main memory in kilobytes.

`cach` cache size in kilobytes.

`chmin` minimum number of channels.

`chmax` maximum number of channels.

`perf` published performance on a benchmark mix relative to an IBM 370/158-3.

`estperf` estimated performance (by Ein-Dor & Feldmesser).

Source

P. Ein-Dor and J. Feldmesser (1987) Attributes of the performance of central processing units: a relative performance prediction model. *Comm. ACM.* **30**, 308–317.

References

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S.* Fourth edition. Springer.

a. Welke kolommen dien je te verwijderen om een Principale-Componenten Analyse te kunnen uitvoeren? Verwijder eventuele kolommen met de verkeerde meetschaal, kolommen die niet zinvol zijn,….
b. Bekijk de correlaties tussen de variabelen. Zijn de gegevens bruikbaar om er een Principale-Componenten Analyse op toe te passen?
c. Voer een Principale-Componenten Analyse uit en interpreteer de resultaten. Hoeveel componenten moet je gebruiken om 95% van de variantie te kunnen verklaren?
d. Zijn er een of meerdere variabelen die uitgesproken doorwegen in het bepalen van de eerste hoofdcomponent? Zo ja welke?
e. Maak een biplot