Module 10: Unsupervised learning (Overview/quizz lecture)

TMA4268 Statistical Learning V2023

Stefanie Muff, Department of Mathematical Sciences, NTNU

March 23, 2023

US Arrest Example

```
##             Murder Assault UrbanPop Rape
## Alabama       13.2     236       58 21.2
## Alaska        10.0     263       48 44.5
## Arizona        8.1     294       80 31.0
## Arkansas       8.8     190       50 19.5
## California     9.0     276       91 40.6
## Colorado       7.9     204       78 38.7
```

Scales:

- Number of occurrence per 100 000 people
- Percentage

US Arrest Example

```
##             Murder Assault UrbanPop   Rape
## Alabama     0.0132   0.236       58 0.0212
## Alaska      0.0100   0.263       48 0.0445
## Arizona     0.0081   0.294       80 0.0310
## Arkansas    0.0088   0.190       50 0.0195
## California 0.0090   0.276       91 0.0406
## Colorado    0.0079   0.204       78 0.0387
```

Scales:

- Number of occurrence per 100 people
- Percentage

Scales:

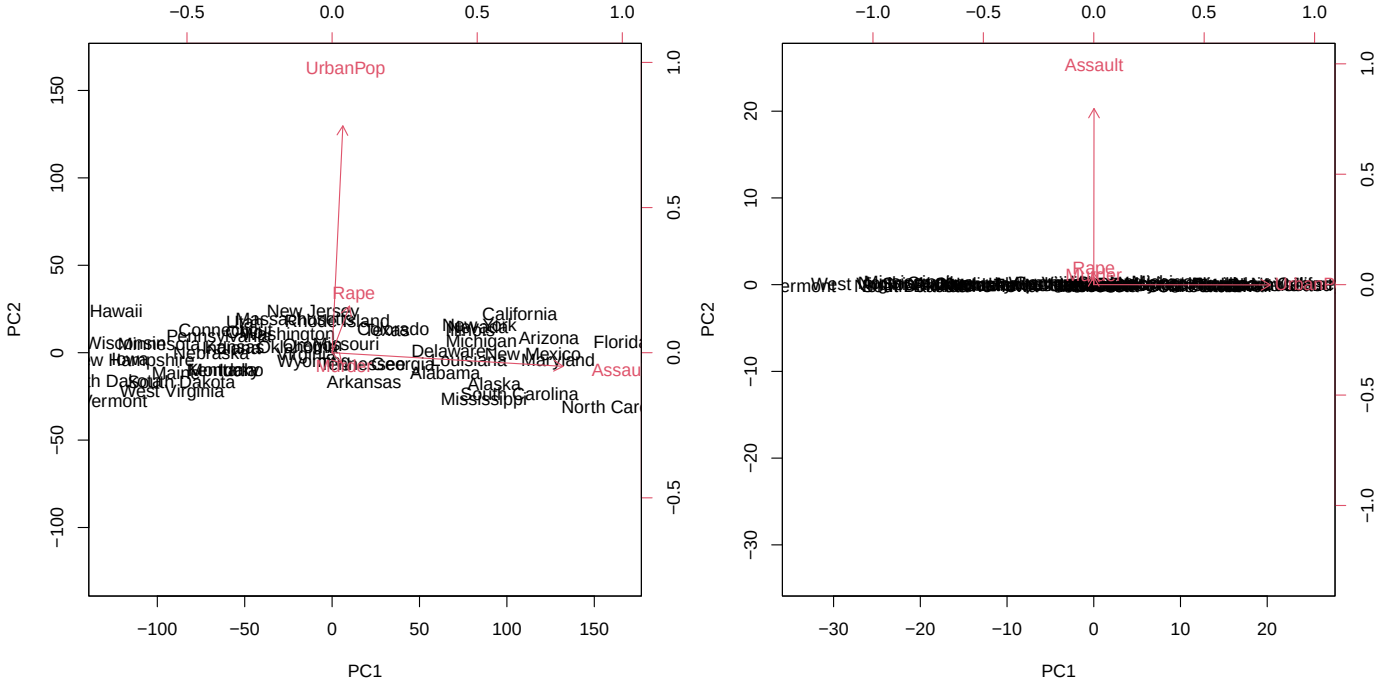
- Number of occurrence per 100 people
- Percentage

PC loadings vectors Φ

	PC1	PC2	PC3	PC4
Murder	0.0417	-0.0448	0.0799	-0.9949
Assault	0.9952	-0.0588	-0.0676	0.0389
UrbanPop	0.0463	0.9769	-0.2005	-0.0582
Rape	0.0752	0.2007	0.9741	0.0723

	PC1	PC2	PC3	PC4
Murder	0.0000	0.0438	-0.0680	-0.9967
Assault	0.0015	0.9968	0.0704	0.0390
UrbanPop	1.0000	-0.0015	0.0002	-0.0001
Rape	0.0003	0.0676	-0.9952	0.0709

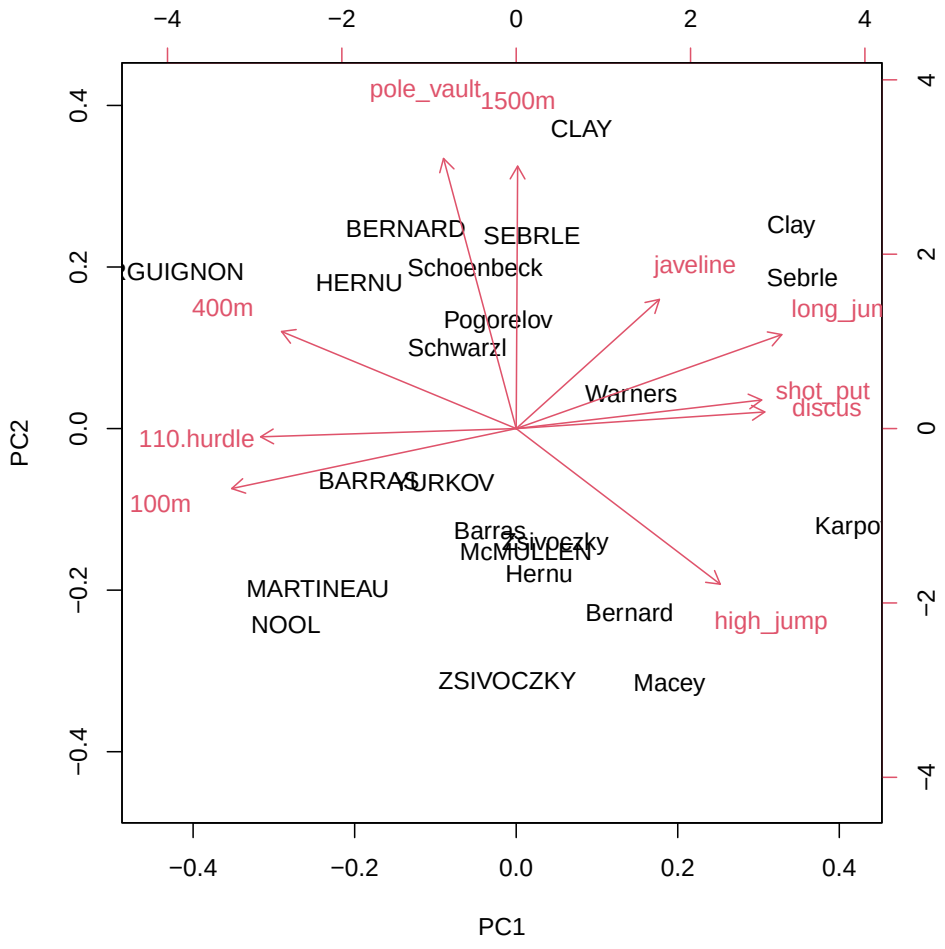
The biplot


UrbanPop
Rape
Hawaii
California
Arizona
Florida
Alabama
Arkansas
Alaska
Mississippi
North Carolina
West Virginia
Assault
PC1
PC2

Example from Compulsory 3, 2020

- We study the `decathlon2` dataset from the `factoextra` package in R, where Athletes' performance during a sporting meeting was recorded.
- We look at 23 athletes and the results from the 10 disciplines in two competitions.

```
##           100m long_jump shot_put high_jump  400m 110.hurdle discus pole_vault
## SEBRLE   11.04      7.58    14.83      2.07 49.81      14.69  43.75       5.02
## BERNARD  11.02      7.23    14.25      1.92 48.93      14.99  40.87       5.32
## YURKOV   11.34      7.09    15.19      2.10 50.42      15.31  46.26       4.72
##          javeline 1500m
## SEBRLE      63.19 291.7
## BERNARD     62.77 280.1
## YURKOV      63.44 276.4
```


pole_vault
1500m
CLAY
BERNARD
SEBRLE
Clay
GUIGNON
HERNU
Schoenbeck
javeline
Sebrle
400m
Pogorelov
long_jum
Schwarzl
Warners
shot_put
discus
110.hurdle
BARRAS
SURKOV
100m
Barras
Karpo
McMULLEN
Hernu
MARTINEAU
NOOL
Bernard
high_jump
ZSIVOCZKY
Macey
PC1
PC2

Scree plot

A graphical description of the **proportion of variance explained (PVE)** by a certain number of PCs (see Fig 12.3 from James et al. (2013)):

Proportion of varianced explained (PVE)

Recap: The PVE by PC m is given by

$$\frac{\sum_{i=1}^{m} z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$$

Clustering

- The aim is to find *clusters* or *subgroups*.
- Clustering looks for homogeneous subgroups in the data.

Difference to PCA?

Clustering

- The aim is to find *clusters* or *subgroups*.
- Clustering looks for homogeneous subgroups in the data.

Difference to PCA?

$\rightarrow$ PCA looks for low-dimensional representation of the data.

K-means vs. hierarchical clustering

See menti.com

K-means clustering

- Fix the number of clusters K.
- Find groups such that the sum of the within-cluster variation is minimized.
- Algorithm?

(Fig 12.8 from course book)

Hierarchical clustering

Bottom-up agglomerative clustering that results in a *dendogram*.

Important in hierarchical clustering

- *Linkage:* Complete, single, average centroid.
- *Dissimilarity measure:* Euclidian distance, correlation. *Other similarity/distance measures?* [1]

[1] Note: Correlation is actually a similarity measure, not a distance measure. Implication?

Hierarchical clustering – example

Note: The representation on the right is not possible in high-dimensional space (i.e., if we have $X_1, X_2, X_3, ..., X_p$).

Hierarchical clustering – example

An exam question from 2022:

Pros and cons of clusterization methods / practical issues

References

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 112. Springer.