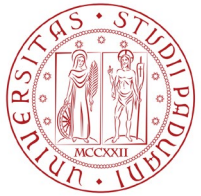


UNIVERSITÀ DEGLI STUDI DI PADOVA

DEPARTMENT OF INDUSTRIAL ENGINEERING

Machine Learning Lesson #7

Academic year 2025-2026

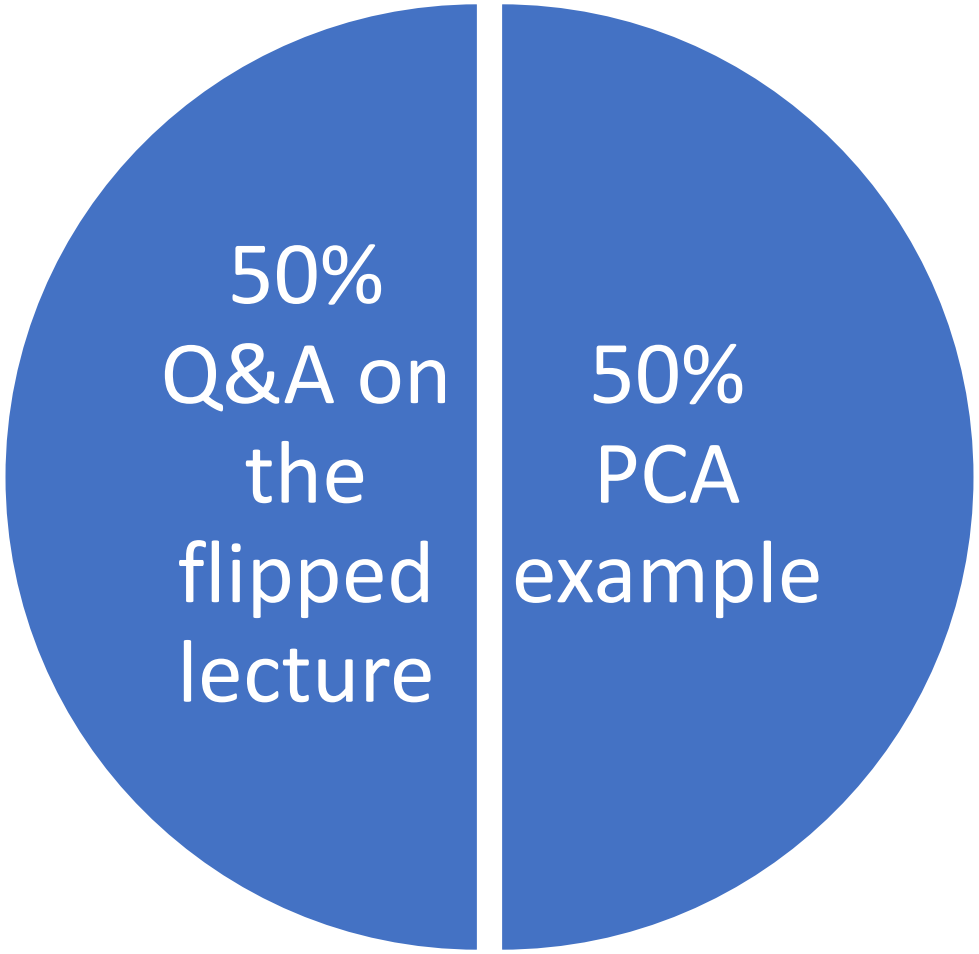
Prof. Pierantonio Facco

CAPE-Lab, Computer-Aided Process Engineering Laboratory

Email: pierantonio.facco@unipd.it

URL: https://research.dii.unipd.it/capelab/

Today's lesson



Q&A on the flipped lecture

White slide because now YOU have to build the lecture with your questions!

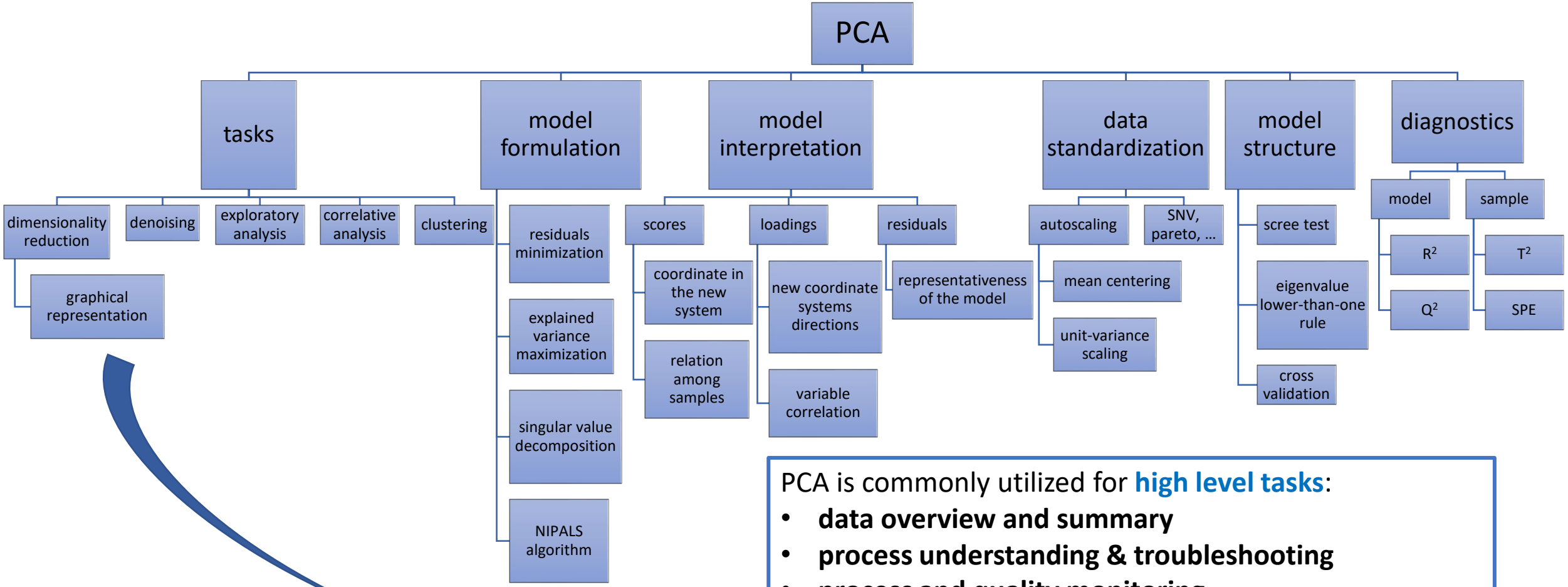
PCA application example

Recap of the last lectures

- When a multivariate dataset $\mathbf{X}$ (i.e., the number of variables $V > 3$) is available we need appropriate methodologies to deal with:
  - dimensionality
  - collinearity
  - noise
  - missing measurements

- Multivariate statistical techniques can face these challenges exploiting:
  - the **correlation structure** of the variable
  - the capability of effectively **summarizing** the information content of data **variability** in few **latent variables**, which are the hidden physical phenomena that drive the system under study and retain all the valuable content of information of the original variables
  - some **projection methods** on reduced spaces (i.e., spaces of reduced dimension) that optimally fit the available data

PCA recap



PCA is commonly utilized for **high level tasks**:

- **data overview and summary**
- **process understanding & troubleshooting**
- **process and quality monitoring**
- **root cause diagnosis of malfunctions**
- **predictive maintenance**

Examples of data overview

PCA in practice

Study of the relation between health and alcohol consumption

Health and alcohol consumption

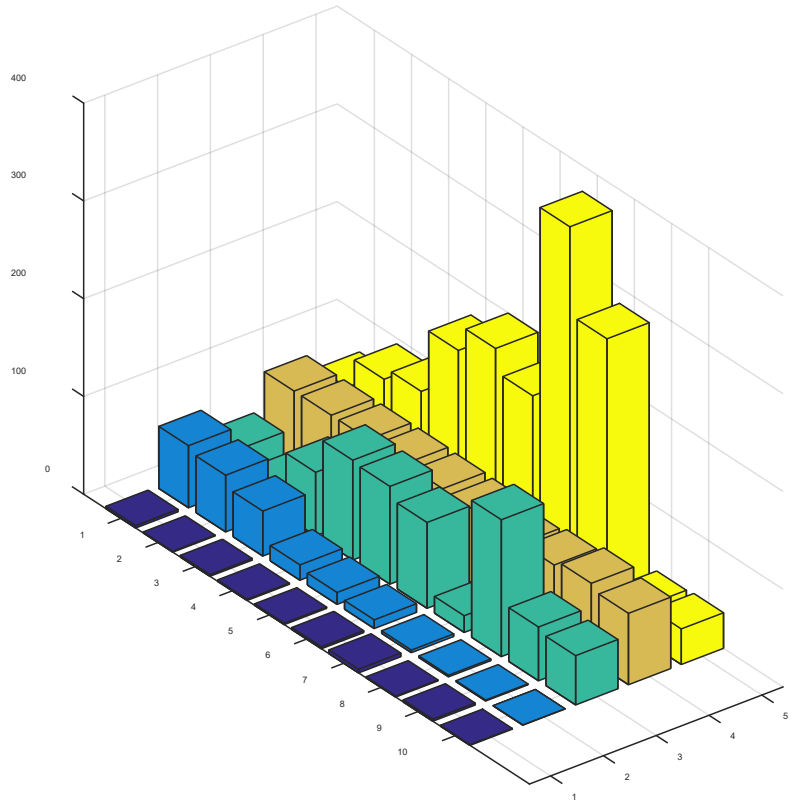
- Available data:
  - 10 observations (on the rows)
    - nations
  - 5 variables (on the columns)
    - wine, beer, liquor consumption and heart disease rate, life expectation

	liquor consumption [L/year]	wine consumption [L/year]	beer consumption [L/year]	life expectancy [years]	heart disease rate [10^5 cases/year]
France	2.5	63.5	40.1	78	61.1
Italy	0.9	58.0	25.1	78	94.1
Switzerland	1.7	46.0	65.0	78	106.4
Australia	1.2	15.7	102.1	78	173.0
Grait britain	1.5	12.2	100.0	77	199.7
USA	2.0	8.9	87.8	76	176.0
Russia	3.8	2.7	17.1	69	373.6
Czech Republic	1.0	1.7	140.0	73	283.7
Japan	2.1	1.0	55.0	79	34.7
Mexico	0.8	0.2	50.4	73	36.4

- **Objective**: **exploratory analysis** of the data

Visual data exploration (1/2)

- Always start data analysis with:
  - **raw data visualization**
  - **system knowledge**



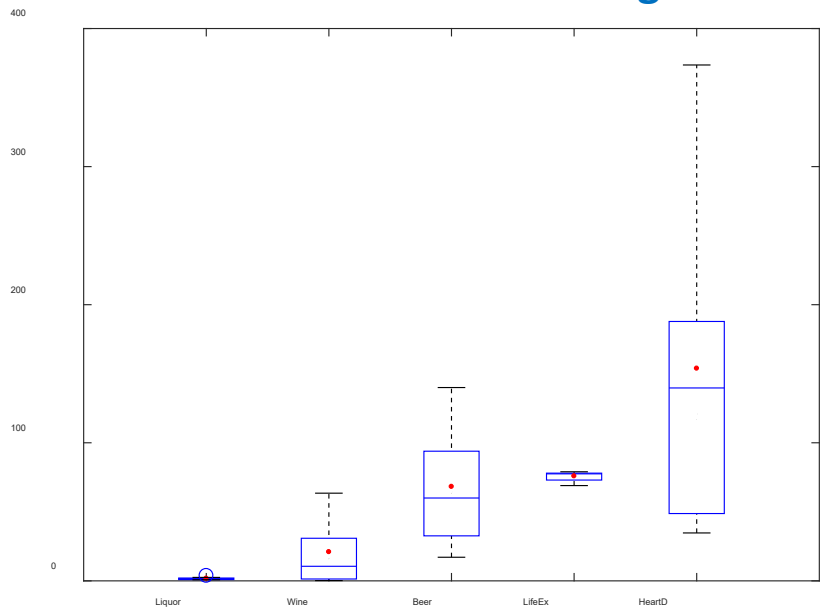
Data have <u>different scales</u>!

<u>AUTOSCALING</u> to give the same weight to all the data (same importance to all the variables)

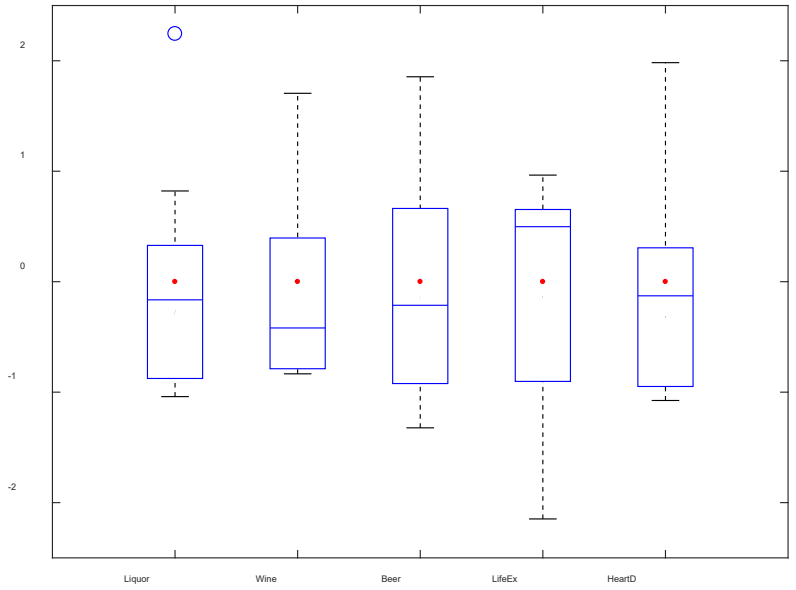
Visual data exploration (2/2)

- Additional suggestions:
  - visualize boxplots for all the variables
  - observe the variables time profiles
  - etc…

before autoscaling



after autoscaling



Model building: eigenvalues and explained variance

- How many selected principal components?
  - scree test
  - eigenvalue-based rule
  - cross-validation

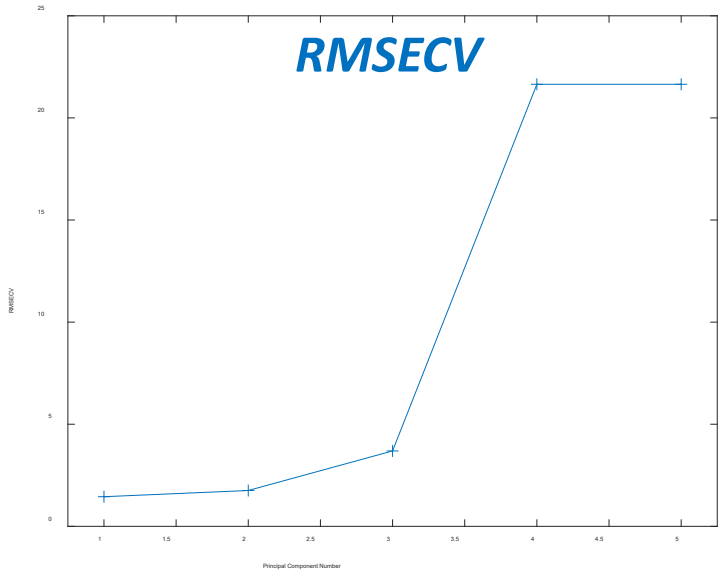
PC	eigenvalue	explained variance (%)	cumulative explained variance (%)
1	2.30	46.03	46.03
2	1.61	32.11	78.14
3	0.58	11.68	89.82
4	0.42	8.44	98.27
5	0.09	1.73	100.00

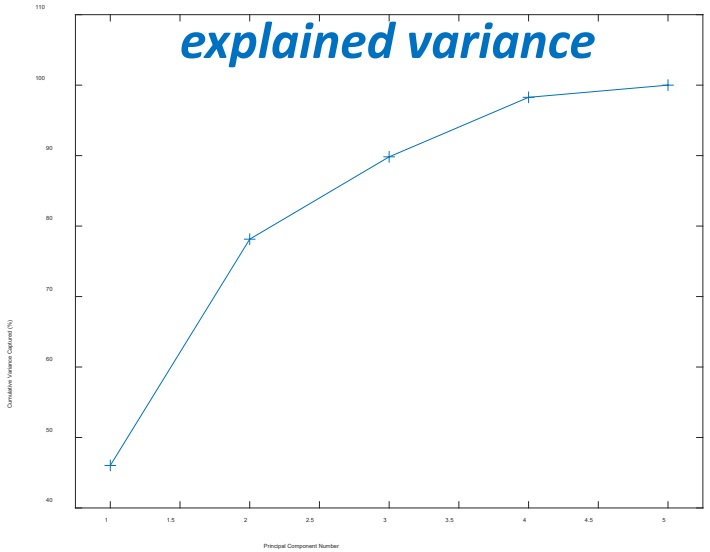
this is the same as the determination coefficient R^2

How to select the correct number of PCs

- Finding the “elbow” on the figure of the explained variance vs. the number of PCs
- Cross validation to find the minimum **RMSECV**

$$RMSECV = \sqrt{\frac{\sum_{n=1}^{N}(x_n - \hat{x}_n)^2}{N}}$$



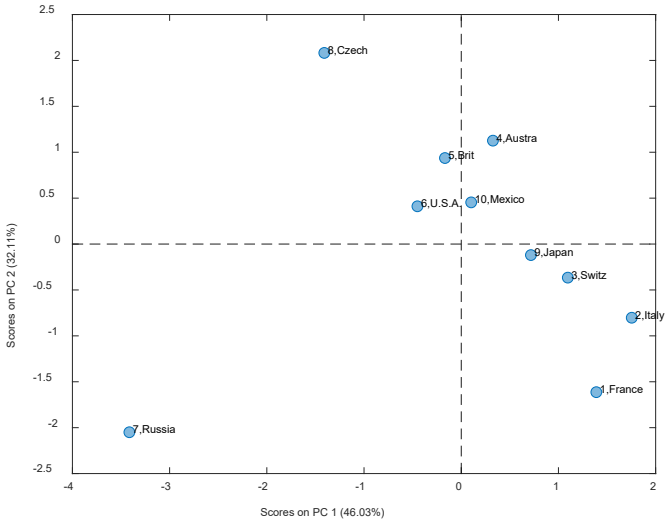


Valle, S., Li, W., Qin, S.J., 1999. Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods. Ind. Eng. Chem. Res. 38

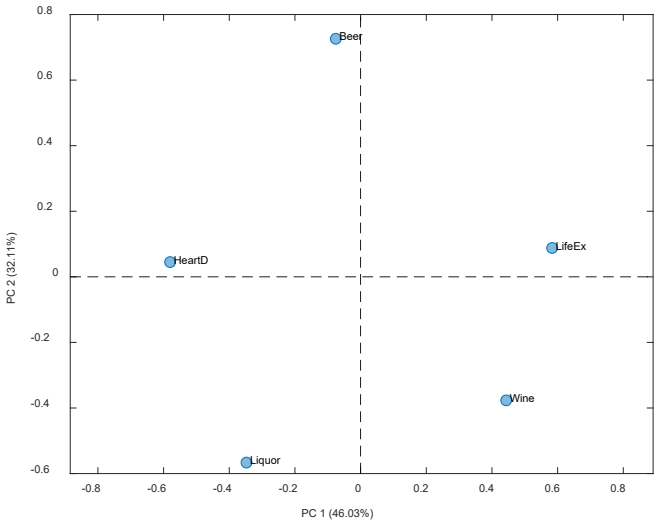
Insight on the PCA model

- Analysis of:
  - scores
  - loadings
  - **residuals**
    - are they low?
    - are they normally distributed?

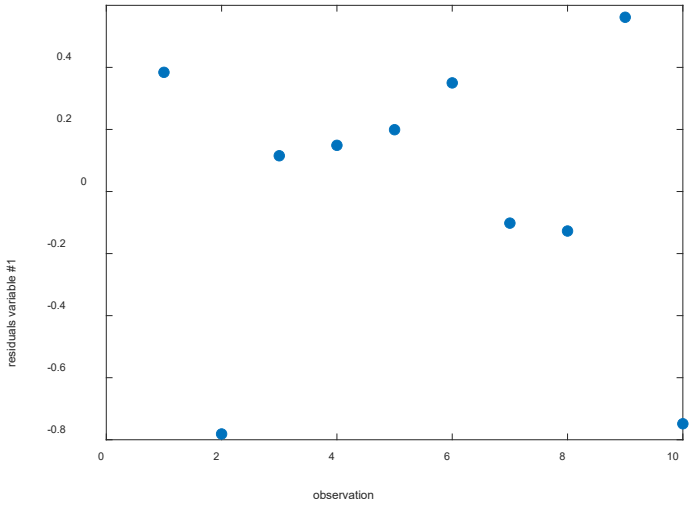
scores



loadings

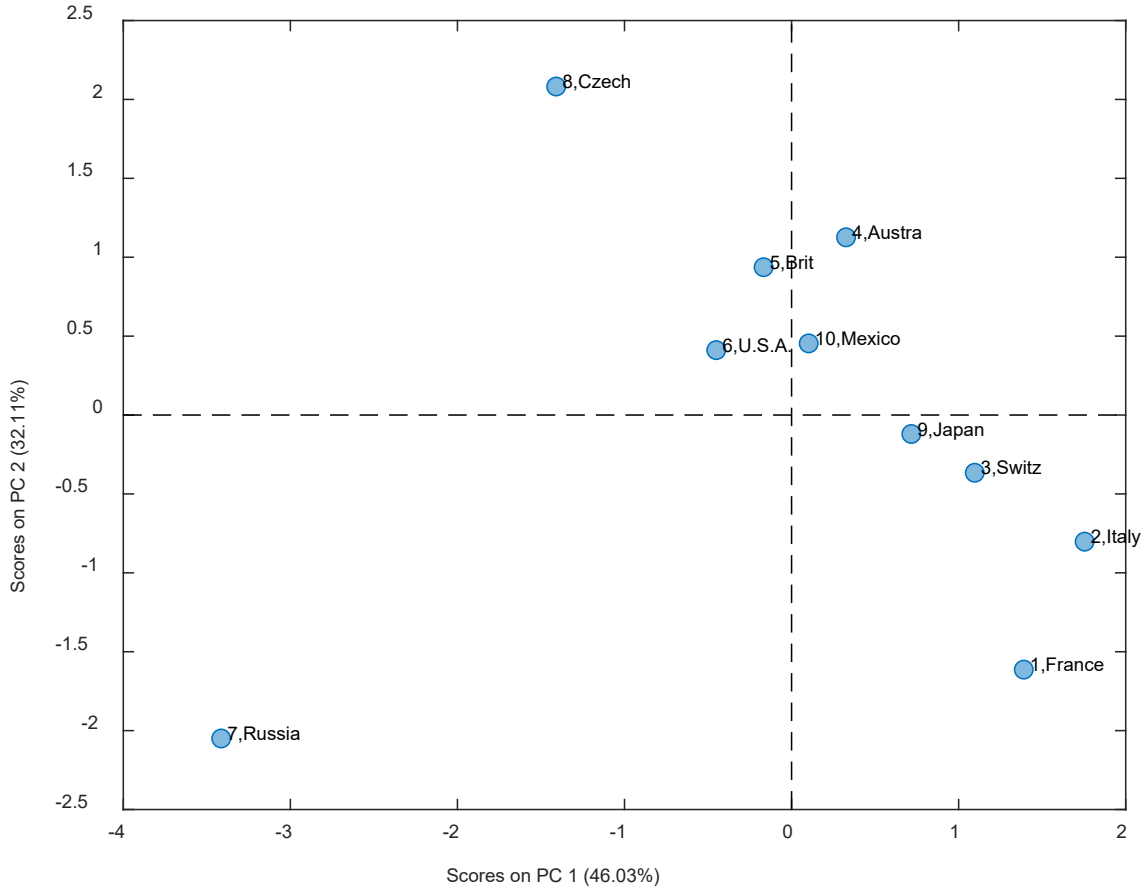


residuals



Relation among observations: score interpretation

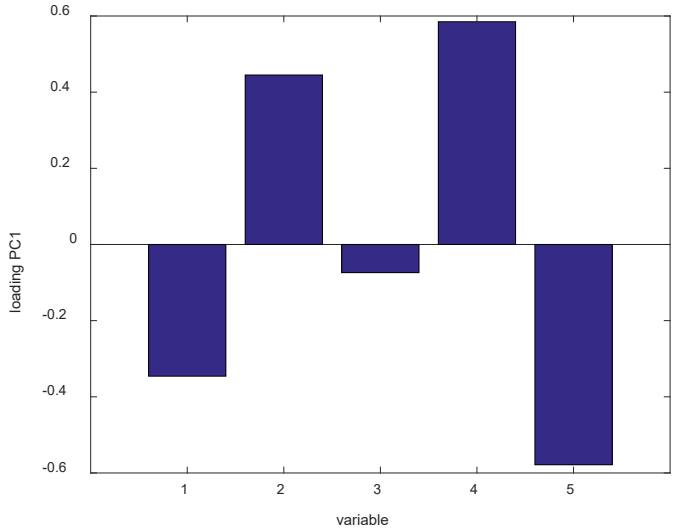
- How are **observations** related?
- Scatter plot of the score space:
  - relation among observations

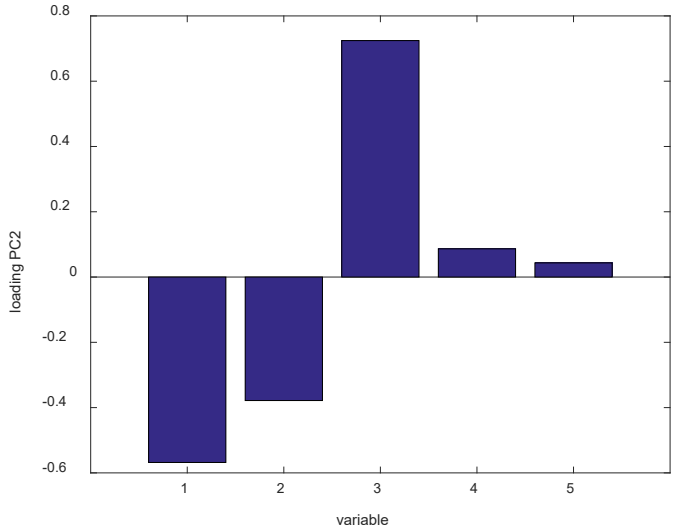


Variables correlation: loadings interpretation

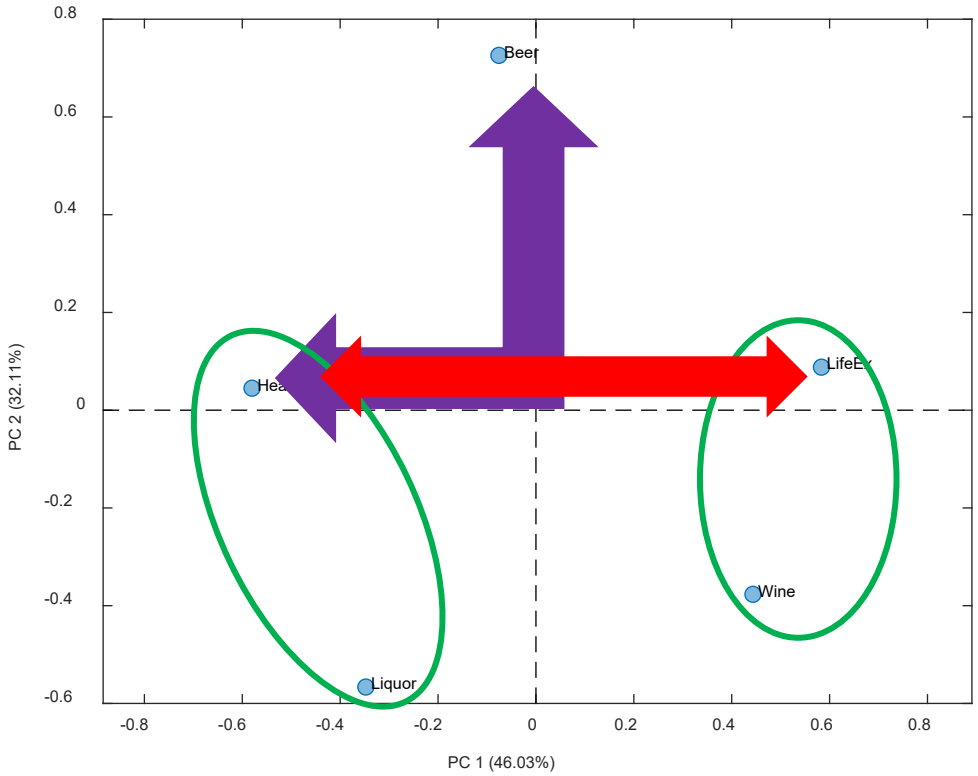
- Two main questions:
  - how are principal components related to the original variables?
  - how are variables correlated?
- Bar plot of the loadings for single PCs:
  - **importance** of the variables on the PCs
    - collinearity of the variables with the PCs
    - most important variables on data variability
  - variables **correlation** and anti-correlation

	variable
1	liquor consumption
2	wine consumption
3	beer consumption
4	life expectation
5	heart disease frequency





Variables correlation



- PC1 (46% explained variance)
  - **positive correlation**
    - wine consumption – life expectation
    - liquor consumption – heart disease frequency
  - **anti-correlation**
    - wine consumption and heart disease frequency
    - wine consumption and liquor consumption
    - life expectation and heart disease frequency
    - life expectation and liquor consumption
- PC2 (32%)
  - beer consumption seems to be independent from life expectation and heart disease frequency
  - beer consumption is partially anti-correlated to wine and liquor consumption
- **Independence** (i.e., orthogonality)
  - beer consumption – life expectation
  - beer consumption – heart disease

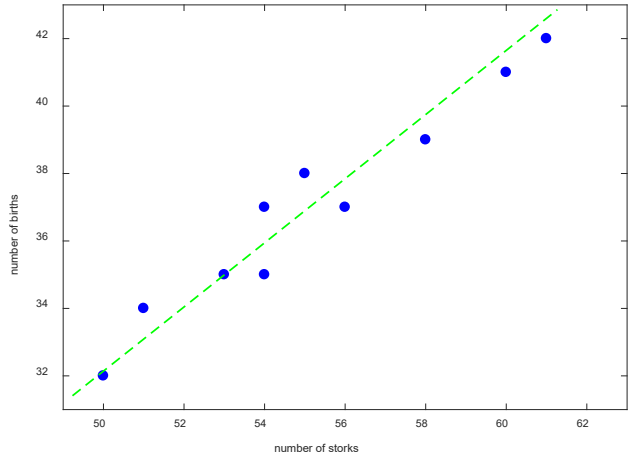
WARNING!!! Causation vs. correlation

- Loadings display **correlations** between variables, and **NOT causation**
  - increasing wine consumption does not increase life expectation!

- For what concerns **causality** conclusions can be obtained discussing with the process/system experts
  - ask to the doctor if wine consumption increases life expectation…

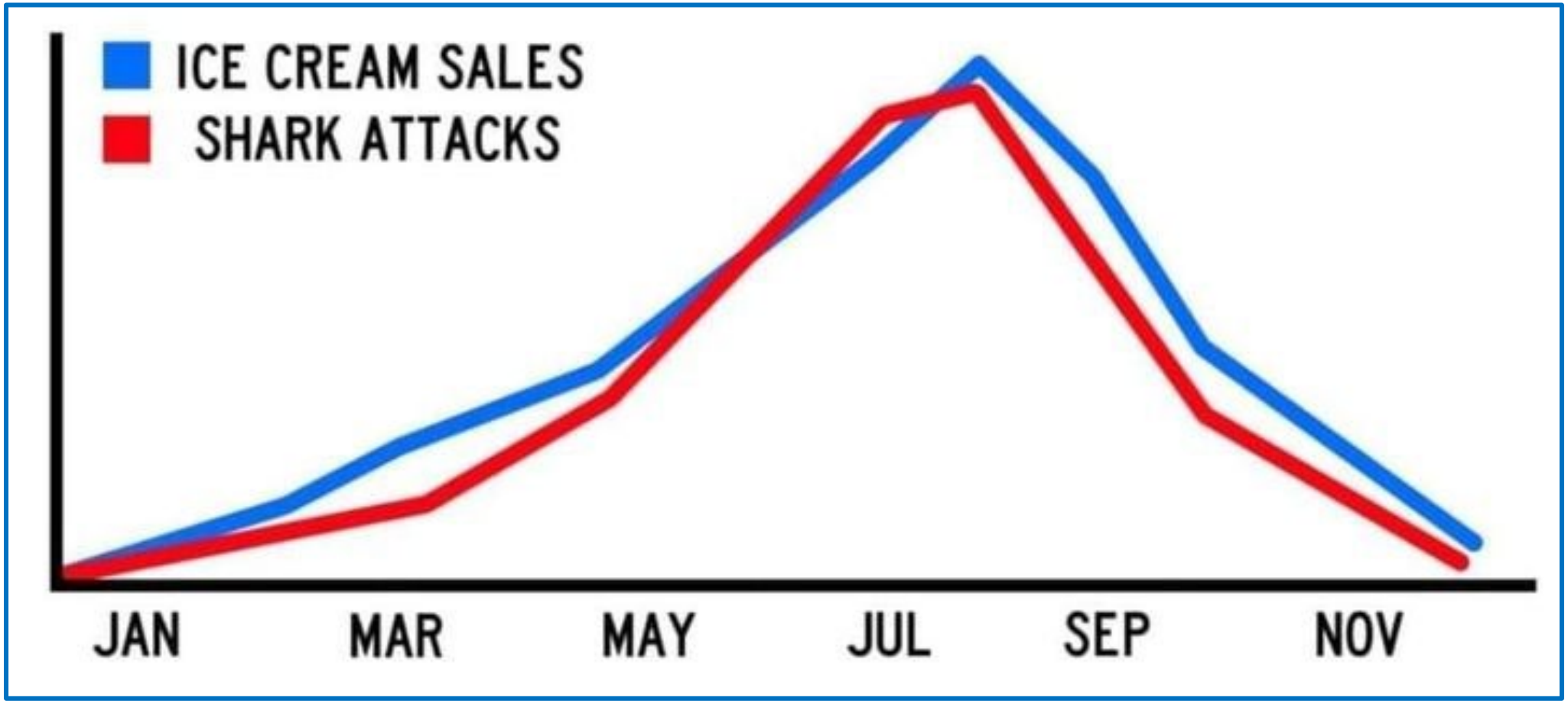
Correlative methods vs. causality

- Multivariate methods are correlative, and not causal!
  - let me tell you a story:
    - in a little Scottish village, there is a lake
    - in the lake there was a strong increase of the number of storks
    - in the same village there was the increase of the number of births



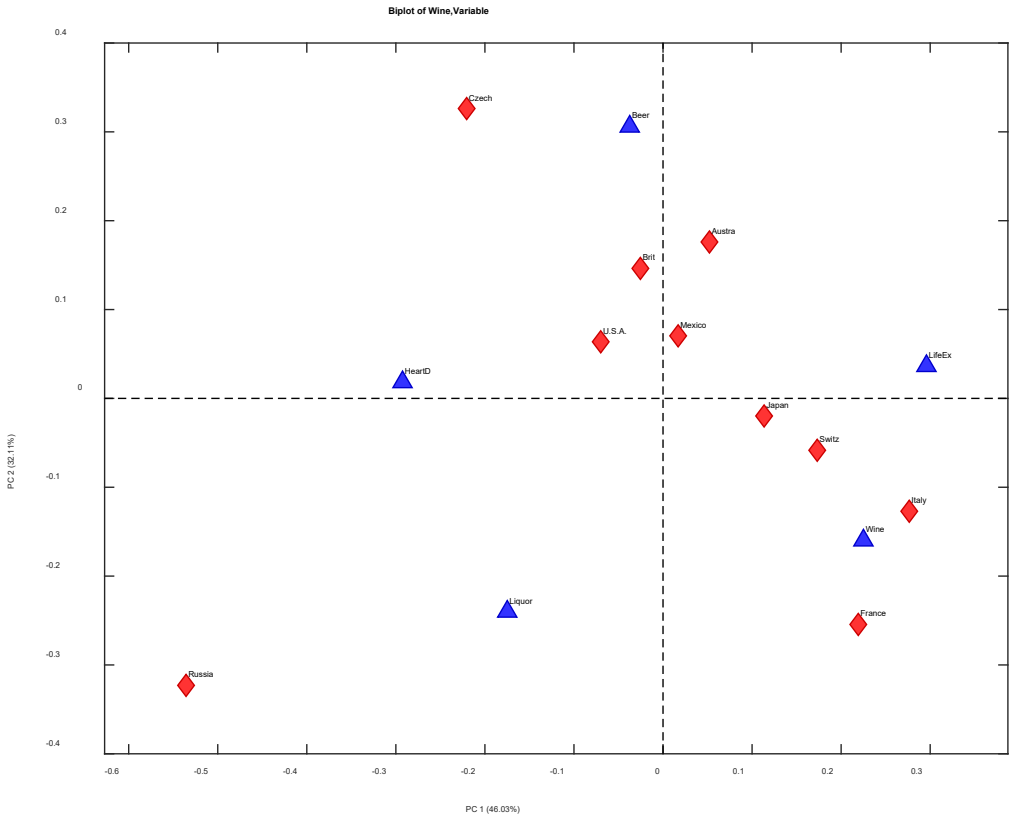
The results of multivariate statistical techniques have to be interpreted in a critical manner!!!

Correlation is not cousation



Biplot: joint score and loading reading

bi-plot = score plot + loading plot



- Japan, Italy, Switzerland and France are the nations with higher life expectation
- France and Italy have the highest wine consumption
- Russia has the lowest life expectation and the highest rate of heart disease and liquor consumption (but low wine and beer consumption)
- Czech Rep. has the highest beer consumption

	liquor consumption	wine consumption	beer consumption	life expectancy	heart disease rate
France	2.5	63.5	40.1	78	61.1
Italy	0.9	58.0	25.1	78	94.1
Switzerland	1.7	46.0	65.0	78	106.4
Australia	1.2	15.7	102.1	78	173.0
Grait britain	1.5	12.2	100.0	77	199.7
USA	2.0	8.9	87.8	76	176.0
Russia	3.8	2.7	17.1	69	373.6
Czech Republic	1.0	1.7	140.0	73	283.7
Japan	2.1	1.0	55.0	79	34.7
Mexico	0.8	0.2	50.4	73	36.4

Take-home message

- Data analysis always start with **data visual inspection** and **system knowledge**
- Data **pretreatment** is one of the key passages to succeed in data analytics
- A golden rule for the **choice of the latent variables number** does not exist:
  - jackknife techniques for cross validation are good and robust tools
  - explain all the variability that is interesting
  - remove all (and only) the noise
- The loadings give clear indications on **variables correlation or independence**
- The scores give information on the **relations between observations**
- The joint reading of scores e loadings (biplot) guarantees:
  - a better **understanding** of the dataset
  - a good summary of the data with a global analysis
- From multivariate statistical techniques you can assess **correlation, not causation**
  - you can draw conclusions about causation only if you have an in-depth knowledge of the system under study

… per sempre a fianco a me!