2.1 Data analytics for dimensionality reduction: Principal Component Analysis (PCA) – Case studies

Prof. Massimiliano Grosso
University of Cagliari, Italy
massimiliano.grosso@dimcm.unica.it
GRICU PhD School 2021
Digitalization Tools for the Chemical and Process Industries
March 12, 2021

1

Outline

- Case study 1
  - Alcohol consumption and quality of life in different countries
- Case study 2
  - Analysis of 1H NMR Spectra of Edible Oils

2

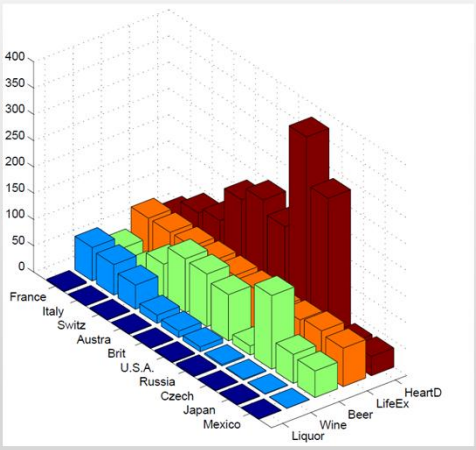
Case study 1

Alcohol consumption and quality of life in different countries

3

Case study 1 – Alcohol consumption vs quality of life

- Beer, wine and liquor consumption (liters per year)
- life expectancy (years) and heart disease rate (cases per 10^5 per year)
  - 5 variables
- Data gathered for 10 countries



4

Case study 1 - Data

variables

	Liquor	Wine	Beer	LifeEx	HeartD
France	2.5	63.5	40.1	78	61.1
Italy	0.9	58	25.1	78	94.1
Switz	1.7	46	65	78	106.4
Austra	1.2	15.7	102.1	78	173
Brit	1.5	12.2	100	77	199.7
U.S.A.	2	8.9	87.8	76	176
Russia	3.8	2.7	17.1	69	373.6
Czech	1	1.7	140	73	283.7
Japan	2.1	1	55	79	34.7
Mexico	0.8	0.2	50.4	73	36.4

samples

5

Case study 1 – Data preprocessing

- Size of variables change from one to another (for example, Liquor values are much smaller than the Beer values)
- This feature, in turn, affects the dispersion (i.e., the variance) of the variables
- A **unity-variance transformation** may be performed to deal with variables characterized by same variance

$$x_{ij} = \frac{x^*_{ij} - \bar{x}^*_{\bullet j}}{s_j} \qquad where \qquad s_j^2 = \frac{1}{n-1}\sum\left(x^*_{ij} - \bar{x}^*_{\bullet j}\right)^2$$

6

Case study 1 – Results

- Eigenvalues of the covariance matrix $X^T X$

Principal comp	Eigenvalue of cov matrix
1	2.30
2	1.61
3	0.584
4	0.422
5	8.64e-2
Total	5

Eigenvalues greater than 1

↓

Variables capture **more** dispersion than the original ones

↓

Included in the PCA model

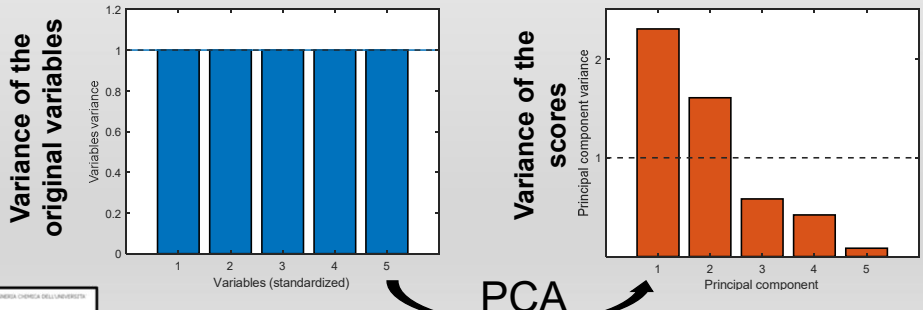
These PCs may be neglected

N.B. Sum of the eigenvalues is equal to the number of principal components as a consequence of the unity variance transformation

7

Case study 1 – Scores variance

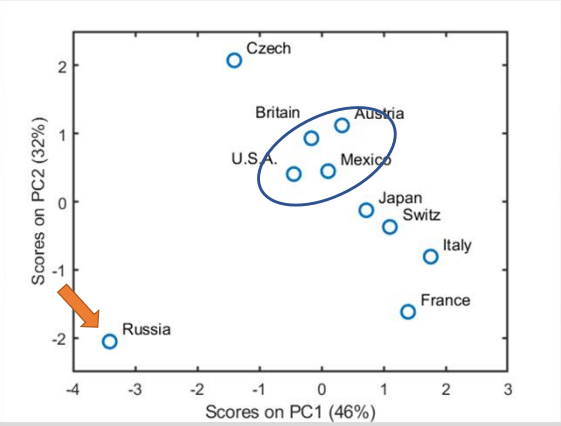
- Original variables vs PCA projection.
- The **first two** principal components explain **more variance** than the original ones. A PCA model with 2 principal components is fair.



8

Case study 1 – Scores analysis

- Countries with similar scores should be similar
  - Russia is quite far away from the other countries
  - U.S.A, Britain, Austria and Mexico are close together
  - Mexico deserves some attention!

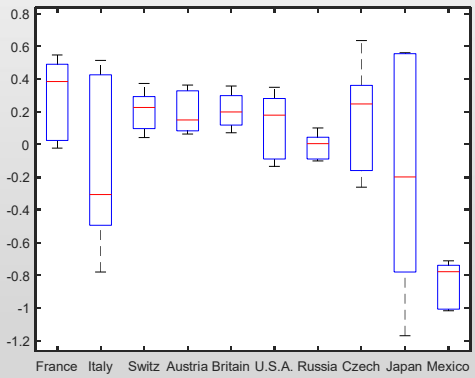


9

Case study 1 – Analysis of residuals

- A rapid inspection shows that the residuals related to the **Mexico** are significantly **larger** than the other ones

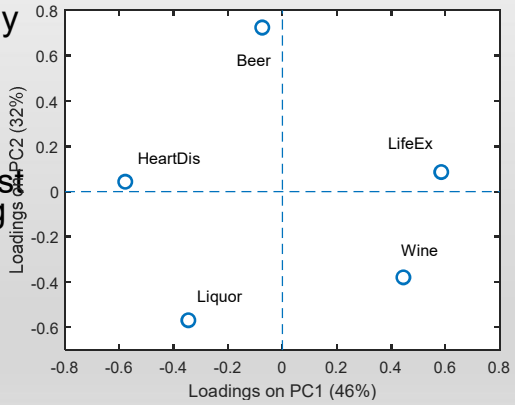
- **Mexico is not well captured by the PCA model**



10

Case study 1 – Loadings analysis

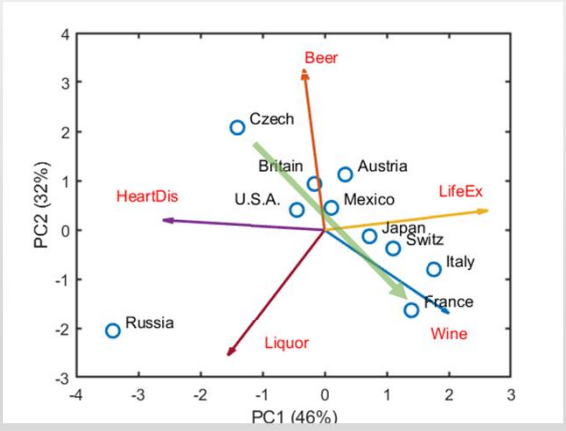
- Variables are far from each other
- Variables which load most significantly into the first PC:
  - life expectancy
  - heart disease rate
    - They are anti-correlated.
- Wine has a positive loading on the first PC, and liquor has a negative loading
- Wine is positively correlated with life expectancy
- Liquor is positively correlated with heart disease



11

Case study 1 – Biplot representation

- **Joint representation** of scores and loadings
- They allow a more detailed analysis
- Example: the “trend” from the Czech Republic to France.
- As a sample moves from upper left to lower right:
  - heart disease and beer consumption decrease
  - life expectancy and wine consumption increase.



12

Case study 2

Analysis of 1H NMR Spectra of Edible Oils

13

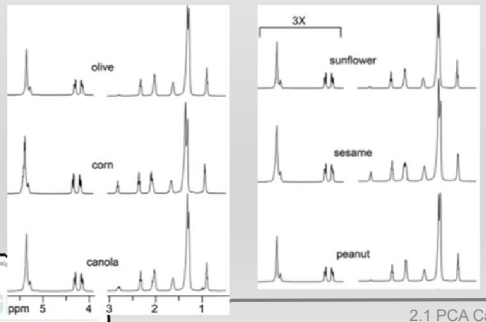
Statement of the problem

- NMR spectra are collected for 6 oils on a 400 MHz spectrometer
  - Canola (5 samples)
  - Corn (5 samples)
  - Olive (5 samples)
  - Peanut (5 samples)
  - Sesame (5 samples)
  - Sunflower oil (3 samples)
- The goal is to analyze them by means of PCA

14

Data

- NMR spectra consist of thousand of variables (chemical shifts)
- In the present study, each sample consists of 1100 chemical shifts



- Differences among spectra cannot be detected by visual inspection

15

Data arrangement

Peculiarity: More variables than observations

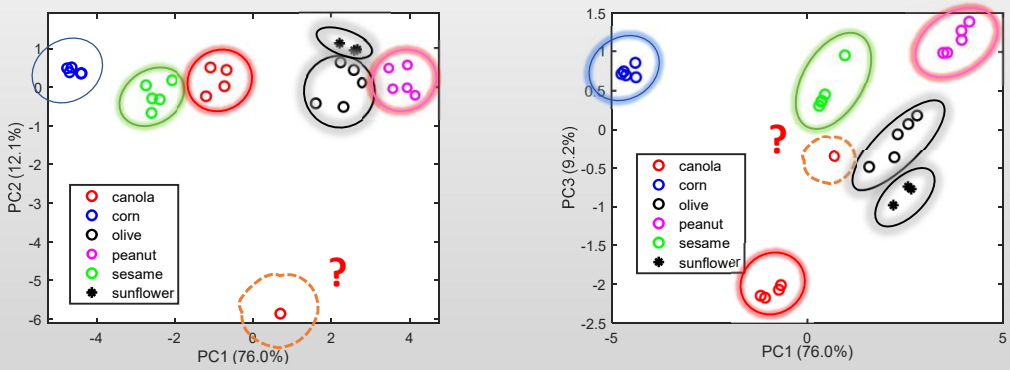
- The data is a (28×1100) matrix
- Unity variance is **not** recommended for this case
  - In general, spectroscopic data are not pretreated with a unity variance scaling
  - A possible alternative is the **Pareto scaling**, used especially with NMR data

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_{\bullet j}}{\sqrt{s_j}}$$

  - However, **no** preprocessing was carried out for the present data

16

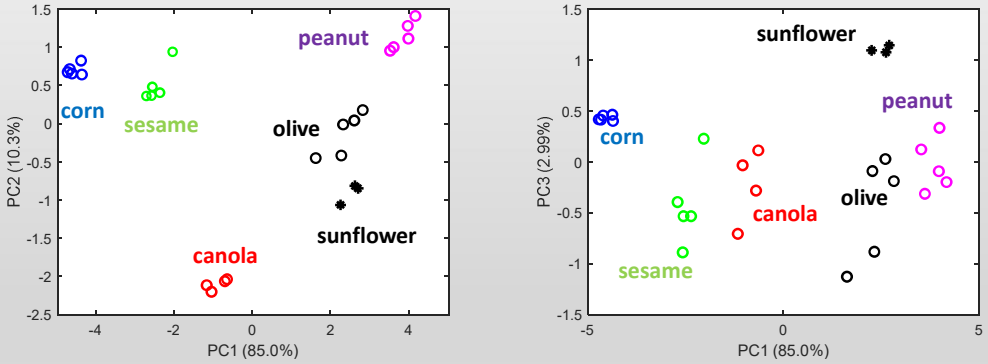
Scores representation



This point is quite far from the canola sample

17

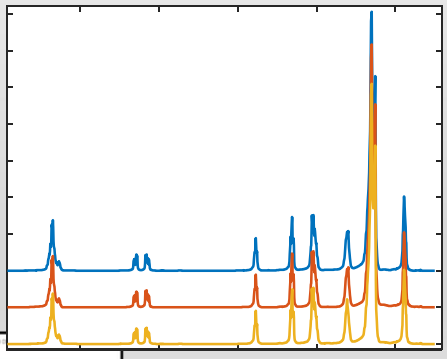
Scores representation – outlier removal



18

Challenge – Classification of unknown spectra

- Three further spectra are then collected
- No knowledge about their provenience



- **Goal**
- To establish, whether possible, which class of oil is «closer» to these ones of unknown origin

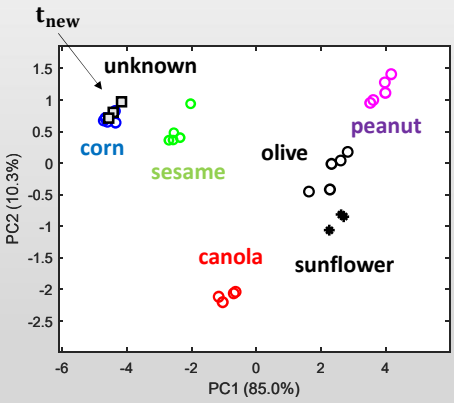
19

Projection of new data on the PCA model

- After mean centering, the spectra can be projected onto the PCA model and compute the related scores
- For each unknown spectra $\mathbf{x}_{new}$:

$$\underset{(1\times 2)}{\mathbf{t}_{new}} = \underset{(1\times 1200)}{\mathbf{x}_{new}} \cdot \underset{(1200\times 2)}{\mathbf{P}_A}$$

- The unknown spectra likely belong to the class of the corn oil



20

References

1. Anderson S.L., Rovnyak D., Strein T.G., 2017, Identification of Edible Oils by Principal Component Analysis of ^{1}H NMR Spectra, *J. Chem. Educ.* 94, 1377-1382