Institute of Applied Physics “Nello Carrara”

Introduction to

Principal Component Analysis

Dr. Leonardo Ciaccheri PhD

The Soft-modelling Approach

Classical Modelling

Theoretically derived model (equations) → The model is validated on relatively few samples. → Prediction are drawn from the model.

Soft Modelling

Many potentially influential variables are measured on many samples (pattern). → A model is extracted from data structure, and validated on a new sampling. → Prediction are drawn from the model.

Large quantity of data → Compression is required

Dimensionality reduction

- Often, many measured variables are redundant. This allow us to reduce the dimension of patterns.
- Redundancy can arise from correlation (*more variables carry the same information*) or poor variance (*the variable does not vary or has poor signal-to-noise ratio*).
- Dimensionality reduction can be achieved by variable selection of by feature extraction. The latter is the case of PCA.
- Feature extraction assumes that measured patterns are influenced by few latent variables, which are the physical sources of their variance.
- Original variables are linearly combined with opportune weights in order to get features likely related to such latent variables.

Decomposition of data matrix

Data Matrix

N objects

	1	2	...	M
1	X_{11}	X_{12}	...	X_{1M}
2	X_{21}	X_{22}	...	X_{2M}
	...	...	...	...
N	X_{N1}	X_{N2}	...	X_{NM}

M centered variables

Columns are linearly combined with opportune **loadings**, thus extracting K new variables, or **scores**.

$K \leq \min(N,M)$.

$$\underset{(N \times K)}{S} = \underset{(N \times M)}{X} \; \underset{(M \times K)}{L}$$

S = score matrix
L = loading matrix (*unitary*)

$$\underset{(N \times M)}{X} = \underset{(N \times K)}{S} \; \underset{(K \times M)}{L^T}$$

PCA Decomposition

How PCs are extracted

A PC is an axis in the data space. The **director cosines** of such axis are given by its **loadings** coefficients. The **score** of an object is the **projection** of its pattern along the PC axis.

The PC axes are chosen imposing the following properties :

1. They are orthogonal
2. PC-1 is the axis of maximum variance
3. PC-n has the highest variance in the sub-space orthogonal to PC-1 ... PC-(n-1)

High-order PCs carry **little information** and can be **discarded**. This reduces both dimensionality and noise.

Orthogonality assures that different PCs are related to different latent variables.

$$\underset{(N \times M)}{\mathbf{X}} = \underset{(N \times K^*)}{\mathbf{S}} \underset{(K^* \times M)}{\mathbf{L}^T} + \underset{(N \times M)}{\mathbf{R}}$$

Residuals = the unmodelled part of data matrix. ($K^* < K$)

Explained variance

$$V_T = \sum_{n=1}^{M} \sigma_n^2 = \sum_{n=1}^{K} \lambda_n^2$$

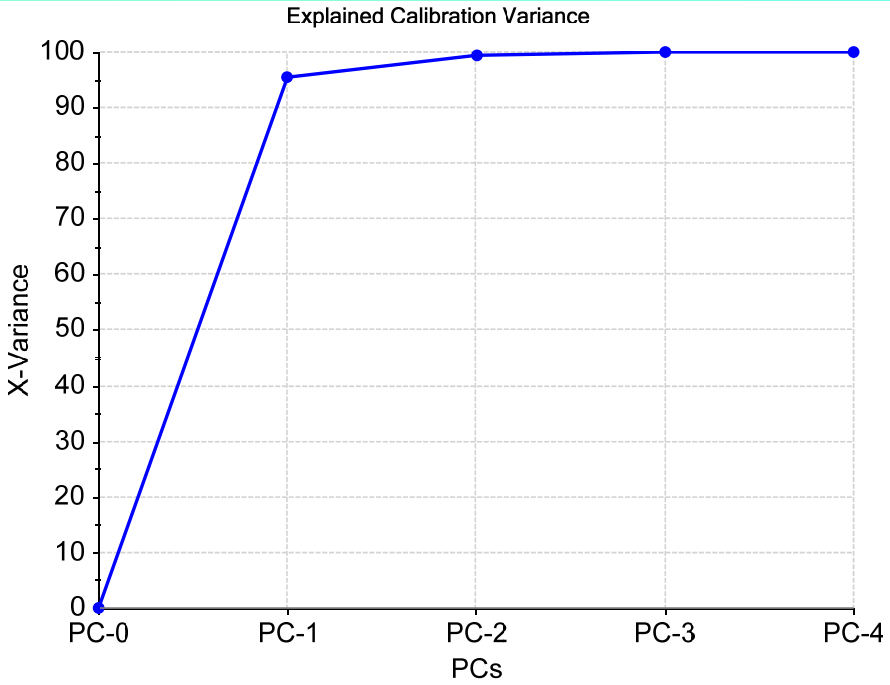
Total variance is invariant between X and S

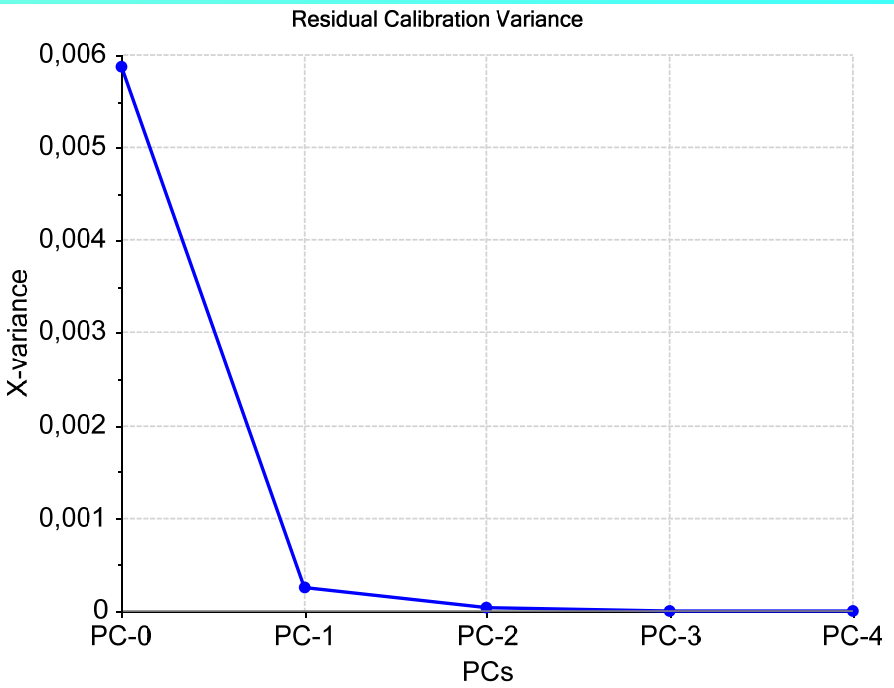
$$EV_{K'} = \sum_{n=1}^{K^*} \lambda_n^2 \Big/ V_T \quad (K^* < K)$$

Explained variance (fraction of V_T)

$$RV_{K'} = \sum_{n=K^*+1}^{K} \lambda_n^2$$

Residual variance (discarded components)





Raman spectra analysis : Score plot

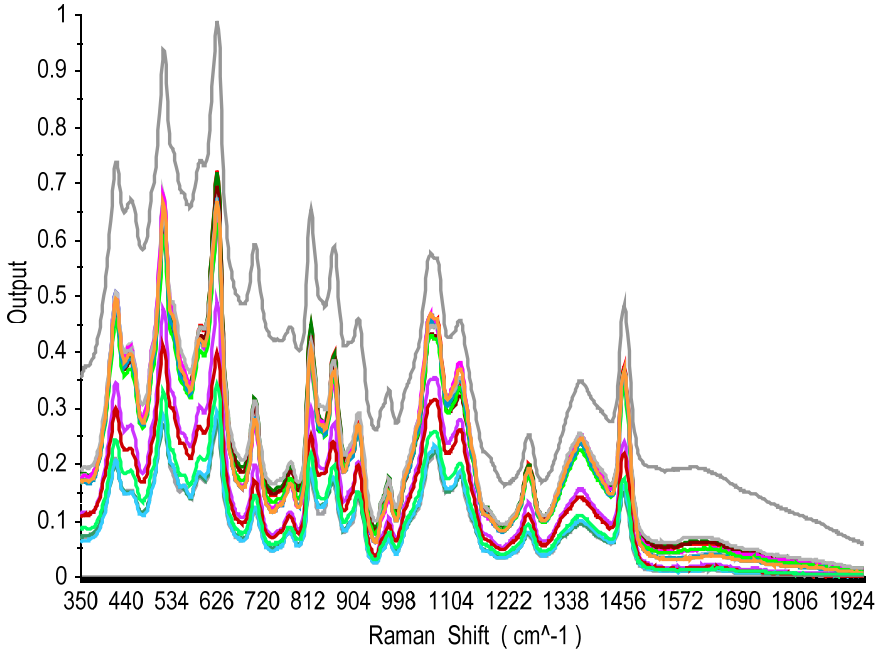
Here it is shown, as example, the PCA of Raman spectra measured on a set of honey samples.

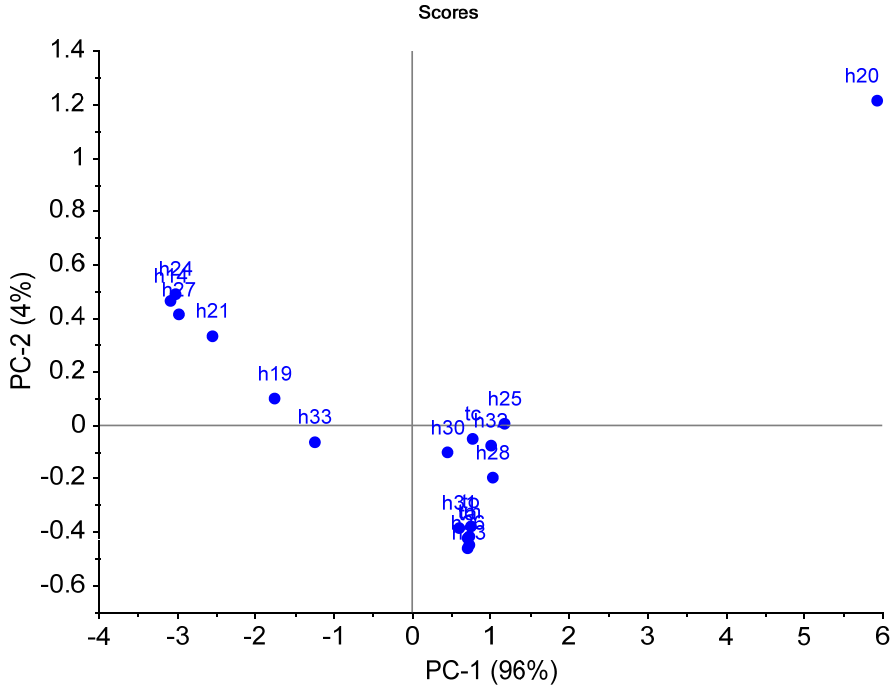
It is **not easy** to detect by naked eye all spectral features and to measure the **dissimilarity** between two spectra.

PCA **scores** allows defining a **distance** between spectra. It reveals at least two **groups** of spectra and a clear **outlier** (h20).

Samples with **average spectra** lie **close to the origin**. Extreme samples lie far from it

Apparently, 2 PCs explain 100% of variance. Indeed, **minor variations** are masked **by the outlier**.

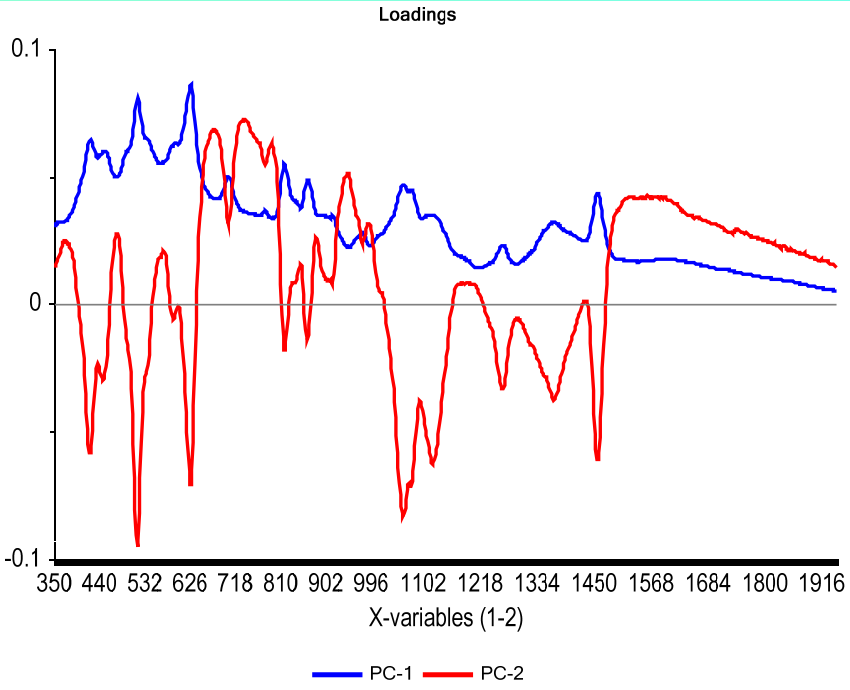




Loading plot: spectral data

Analysis of loading spectra allows detecting **important variables** (*variables with high variance*) and **correlations** among variables. It also allows **interpretation** of scores, explaining why samples are different.

- Variables with loadings far from 0, either positive or negative, are important.
- Loadings with **equal sign** in a spectrum, express a **common mode** variation.
- Loadings with **opposite sign** in a spectrum, express a **contrast** between variables.



PC1 is mainly related with background spectrum.

Loadings are proportional to Raman intensity. It suggests that a multiplicative factor affects spectra.

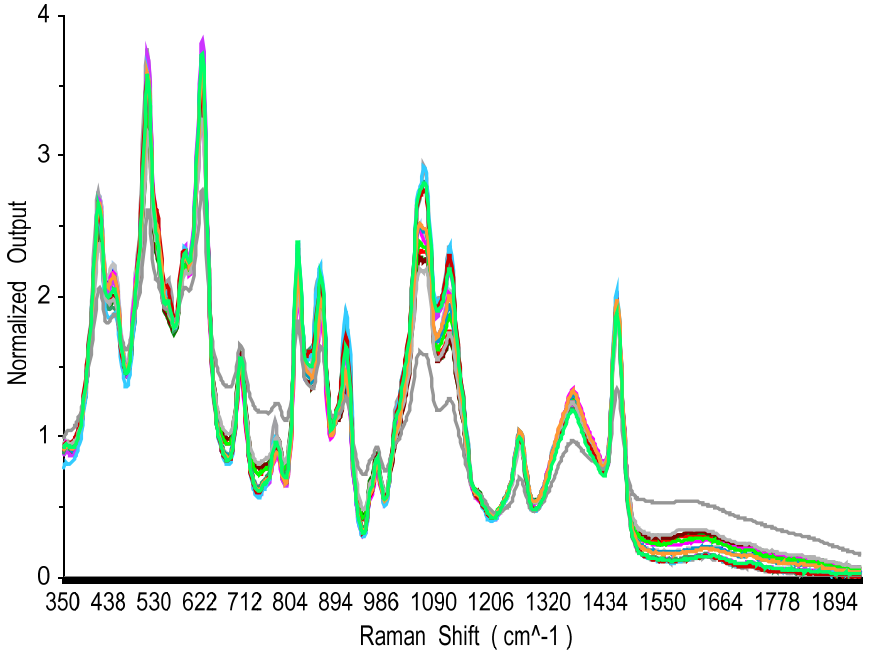
Normalization could be required.

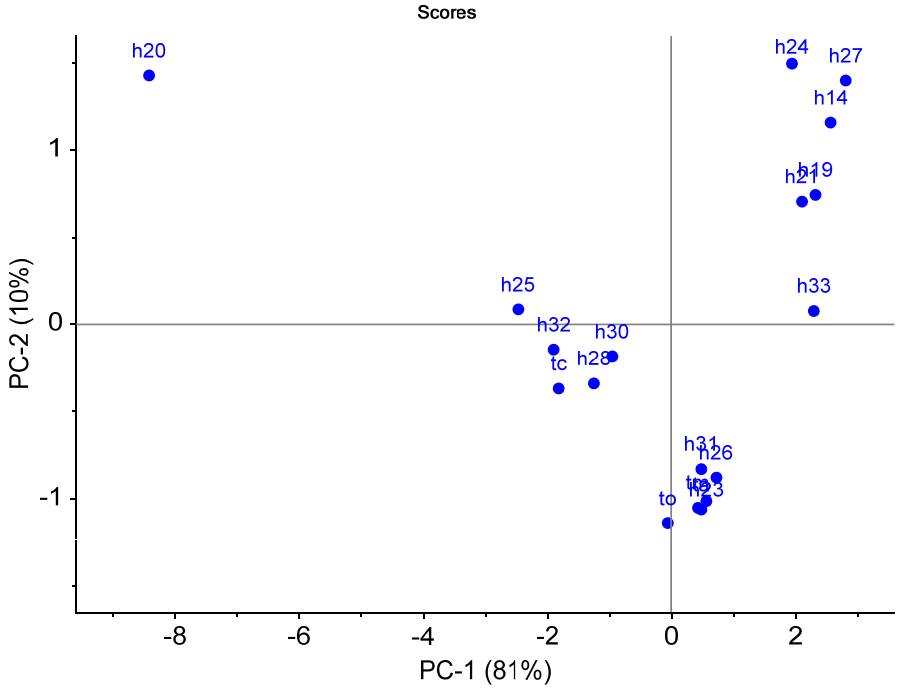
PC2 explains the contrast between the Raman peaks and the background (opposite sign).

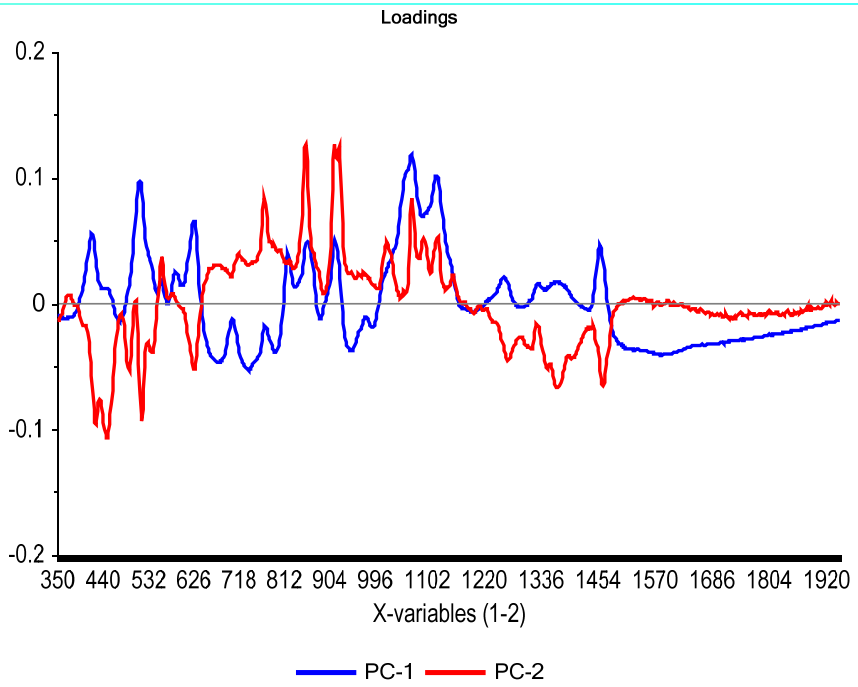
It represents the variation of Raman peaks relatively to background.

Normalized spectra

- Variation of background is strongly reduced
- **Contrasts** between different **Raman peaks** begin to appear in the PC2 loadings.
- Now 2 PCs explains only 91% of variance.
- Sample **h20** is still an **outlier**.







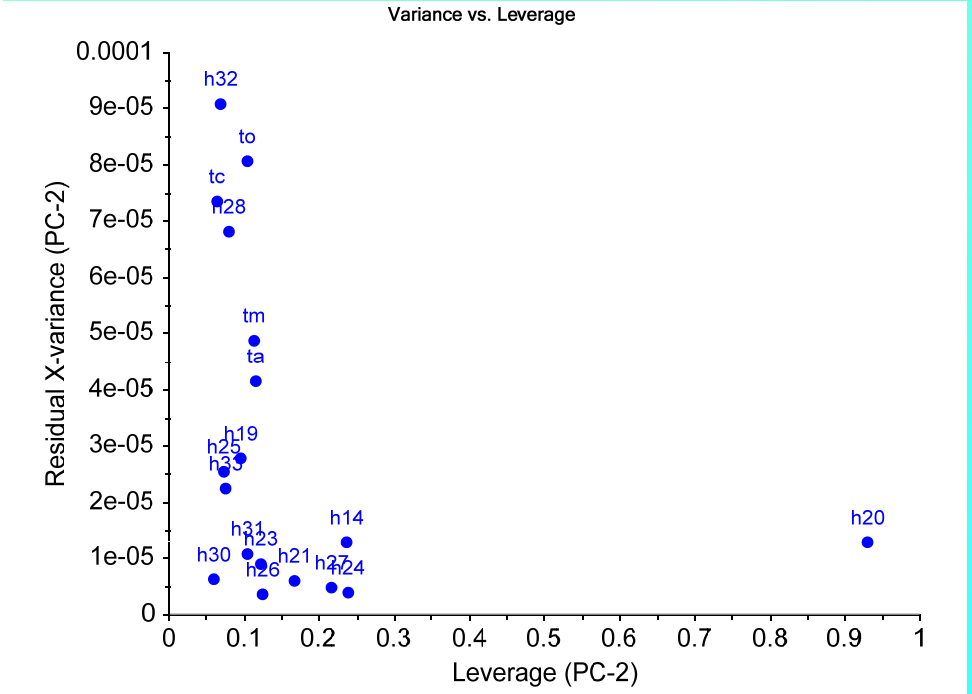
Influence plot

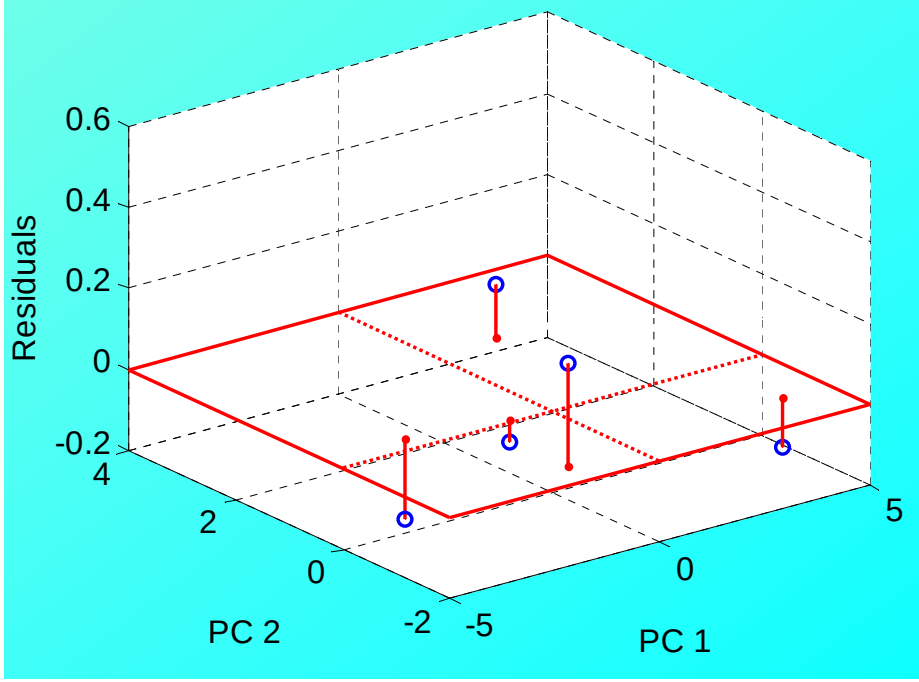
$$\mathrm{Res}_i = \frac{1}{M}\sum_{m=1}^{M} R_{im}^2$$ **Residuals**

$$Lev_i = \frac{1}{N} + \sum_{k=1}^{K} s_{ik}^2 / \lambda_k^2$$ **Leverage**

Residuals variance: mean of squared residuals for each sample. It say how well a sample is described by its projection onto PCA space.

Leverage: weighted distance from the center of the PCA space. It say how much a sample influence PCA model.

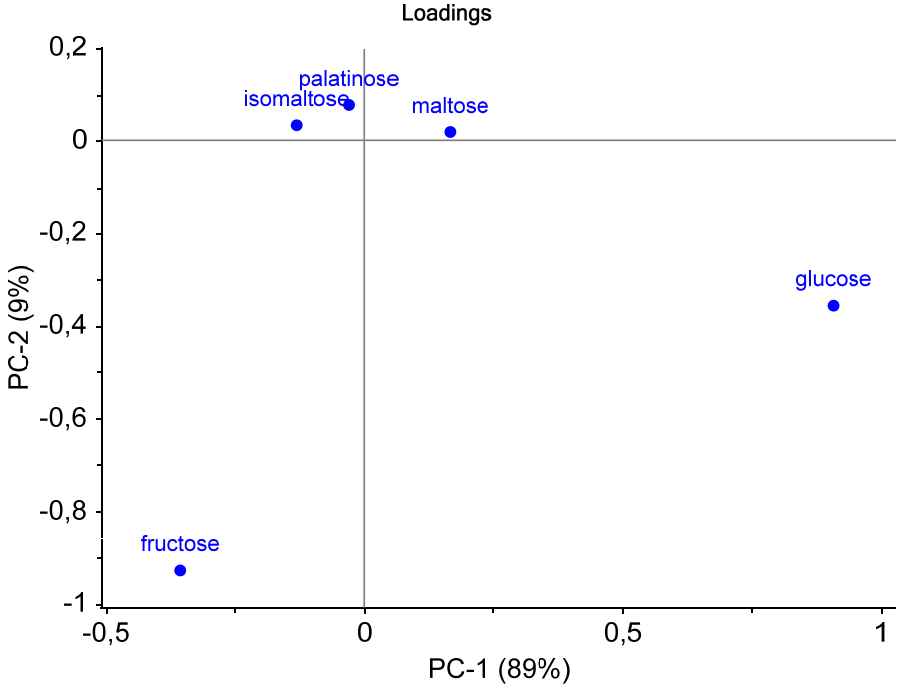


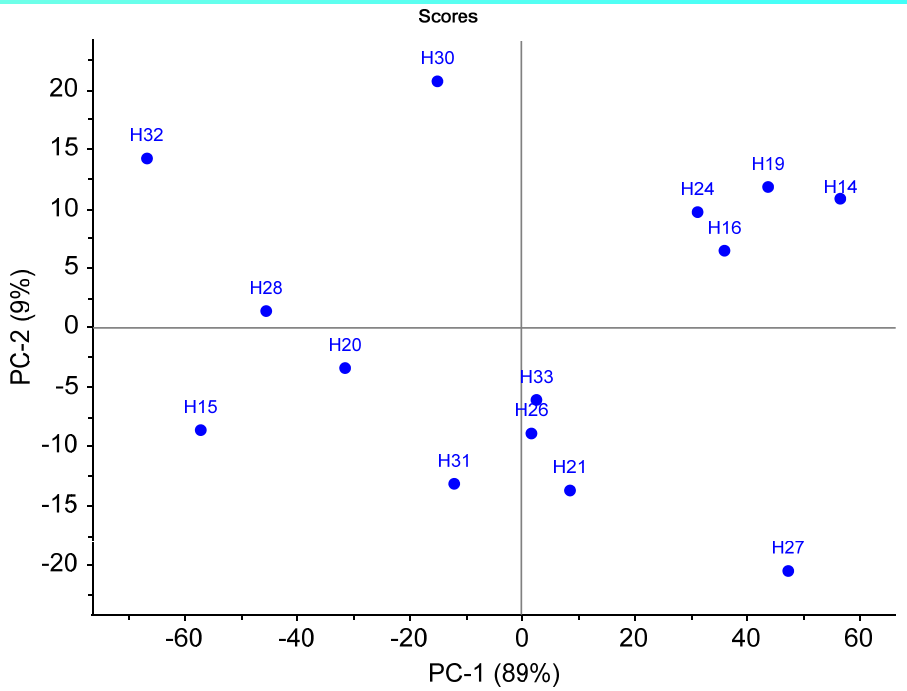


Loading plot: non-spectral data

When there are few variables it is more convenient show loadings as a **scatterplot**, plotting two PCs at a time.

- Points **far from origin** in loading plot represent **important** variables.
- Points with high, absolute, projection along an axis are important for that component.
- Strongly correlated variables have points that are close each other (*direct correlation*) or opposite to origin (*inverse correlation*).
- By comparing score and loading plot, differences between samples can be interpreted.





Application of PCA

Now we will see a case of multivariate data analysis in practice. We will start with an explorative analysis by PCA, and then we will shift to supervised analytical tools for achieving specific goals.

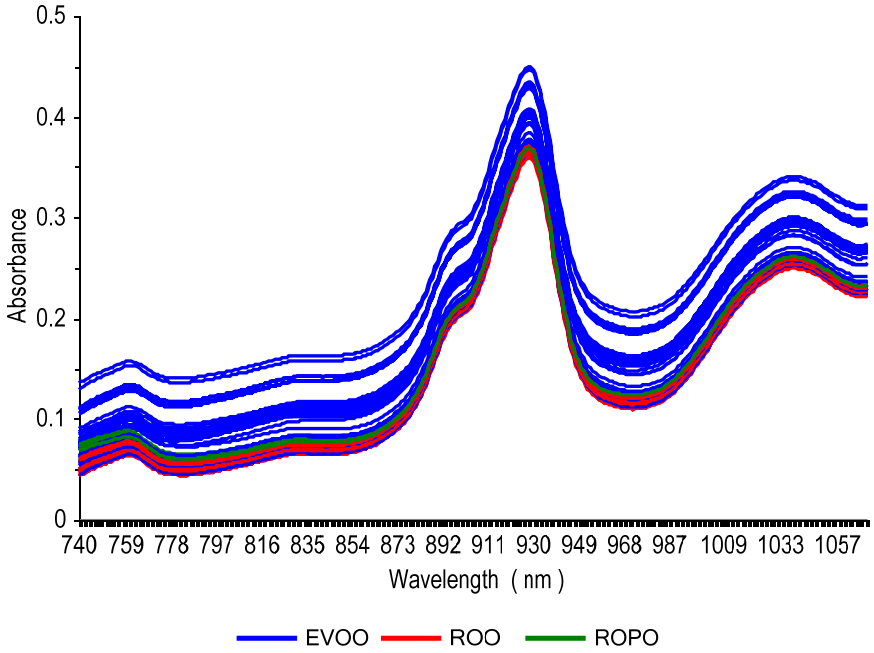
For this purpose, we will introduce a new tool, **Linear Discriminant Analysis,** for sample **classification**.

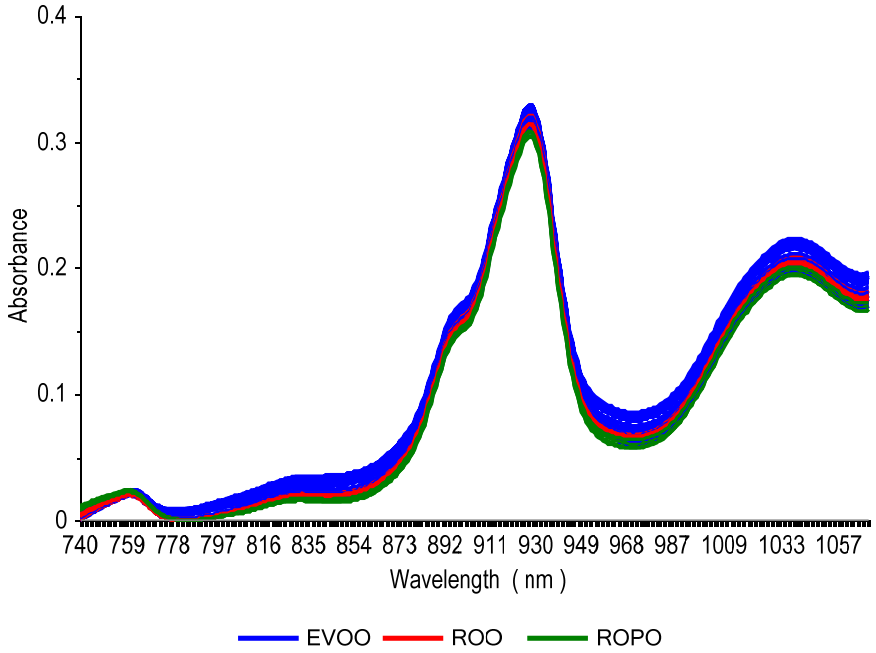
This illustrative data set is made by **Near-Infrared** (NIR) absorption spectra of 60 samples of Olive Oil of three different categories, called EVOO, ROO and ROPO.

- **EVOO** (Extra-Virgin Olive Oil) is high quality olive oil, extracted only with mechanical means, without chemical or thermic treatments.
- **ROO** (Rectified Olive Oil) is low quality olive oil, chemically treated for eliminating defects.
- **ROPO** (Rectified Pomace and Olive Oil) are similar to ROO, but contain also Pomace oil. Pomace is the residual of oil milling, and the oil chemically extracted from it has very low quality.

Visual Inspection of data

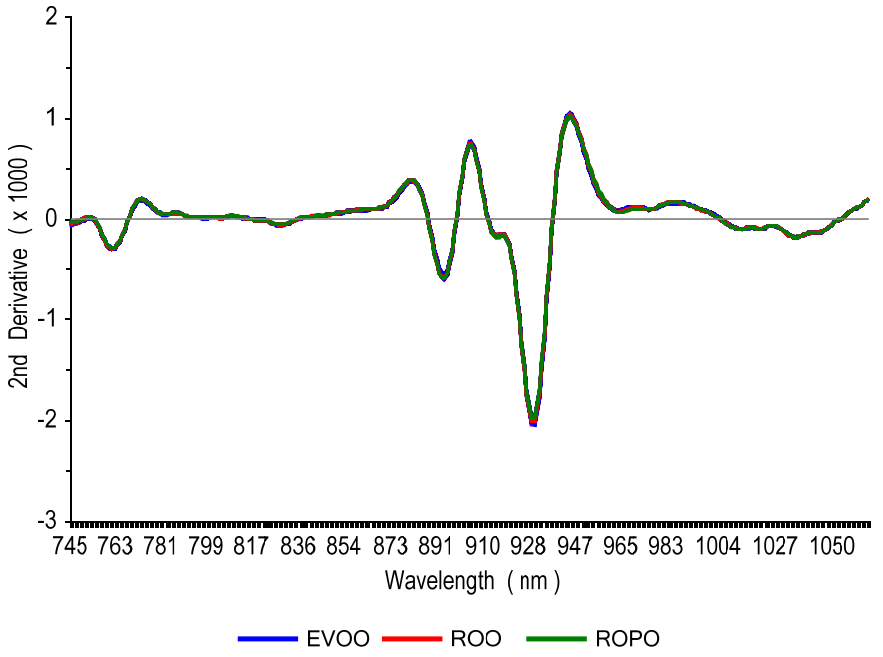
- Visually inspecting spectra before starting analysis is always advisable.
- In this case a strong **baseline shift** is clearly visible.
- Correcting spectra for offset show that a **baseline tilt** is also present.
- This effects arise from **scattering** or instrumental **noise**.
- Both effects can be corrected simultaneously by taking **2nd derivative** of spectra.

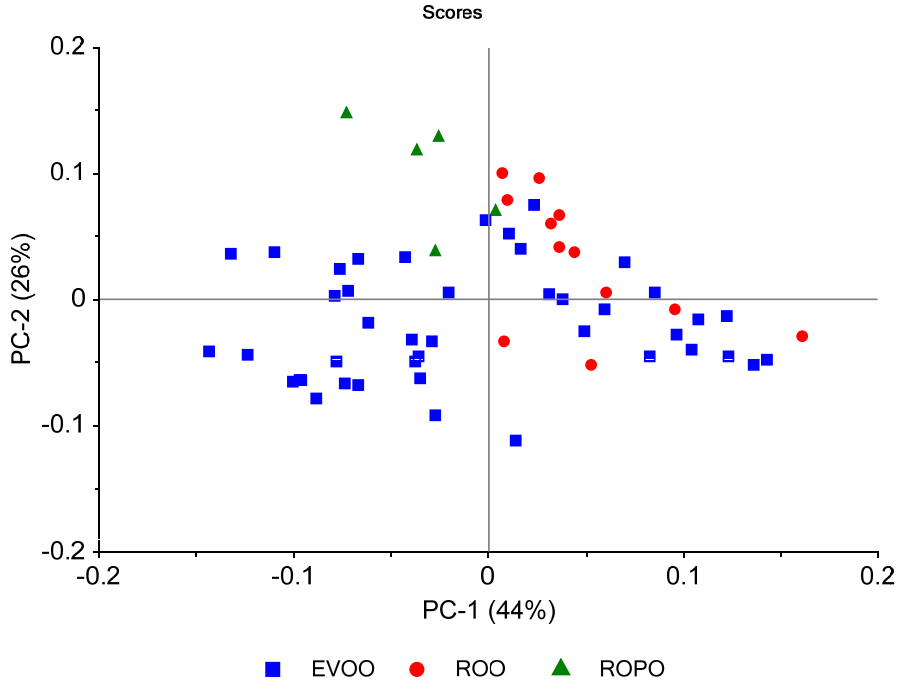


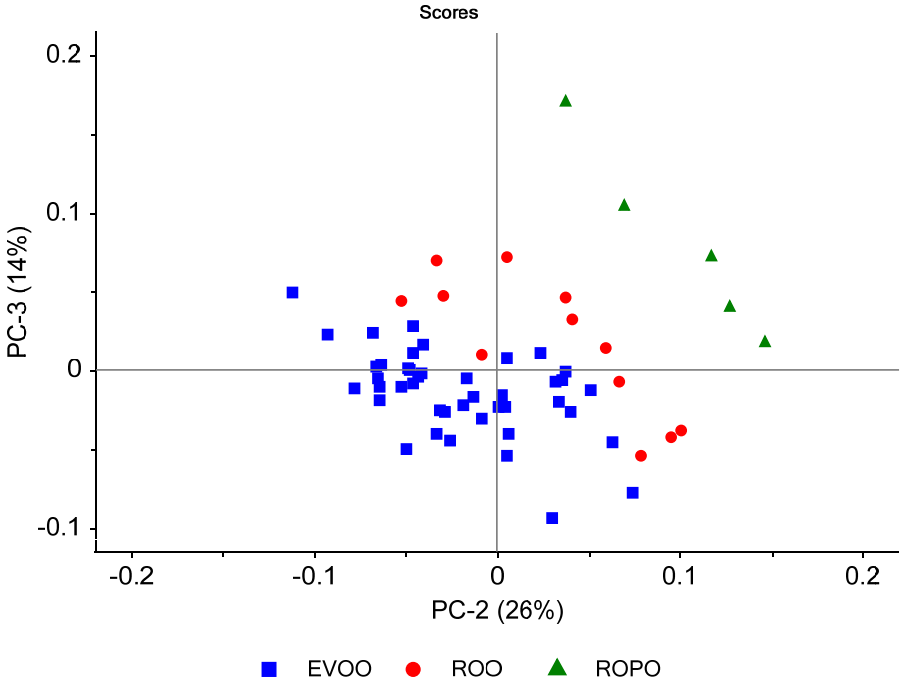


PCA of spectra

- Differences between spectra are weak. PCA is necessary.
- No object grouping is detected along **PC1**.
- **PC2** and **PC3** achieve a rather good splitting of oil categories.
- The three groups are **aligned** among the same direction. this suggest that the same latent variable splits all three groups.







Supervised analysis (LDA)

Supervised analysis is specifically aimed to predicting a qualitative or quantitative target variable. Quantitative prediction will be the argument of next lesson. here we will focus on qualitative analysis.

Dataset Splitting

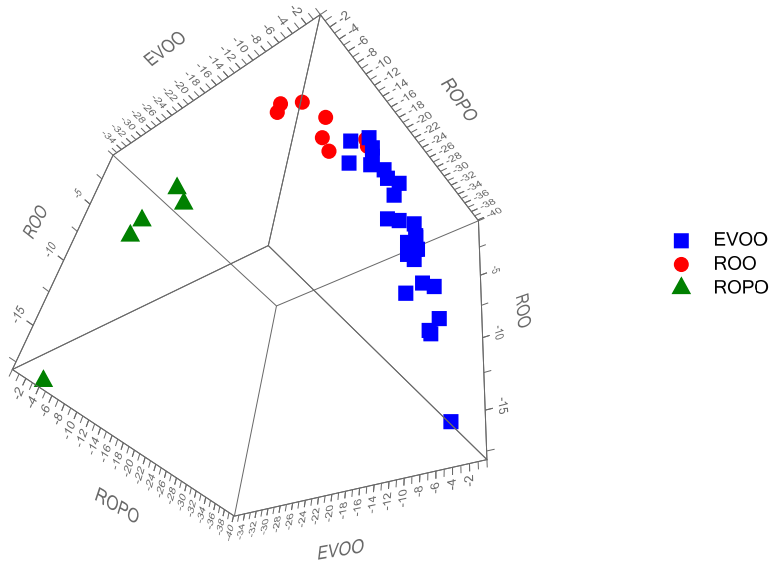
- In order to **avoid overfitting**, it is advisable to test models on validation samples that are not used to train it.
- Dataset has been split in a training set (40 samples) and a validation set (20 samples).

Linear Discriminant Analysis (LDA)

- LDA is classifier that can be used on either PCs or original variables. Because it requires that the number of object in each class exceeds that of variables, PCA is often used to reduce dimensionality, before applying LDA.
- LDA assumes that all classes follow a multivariate Gaussian distribution. On that basis, membership probabilities for all classes are calculated as function of position in PCA space. Each sample is then assigned to the most probable class.

LDA

- LDA generates 3 **scores** (one per class), which are related to the logarithm of **membership probability**. Each sample is assigned to the top-scoring class.
- Classification accuracy is **100%** on training set and **90%** on test set, which is a good result.
- All ROPOs were assigned to training set because they were too few for being split.



Training	Actual		
Predicted	EVOO	ROO	ROPO
EVOO	27	0	0
ROO	0	8	0
ROPO	0	0	5

Test	Actual		
Predicted	EVOO	ROO	ROPO
EVOO	15	1	0
ROO	1	3	0
ROPO	0	0	0

Conclusions

PCA achievements

- Dimensionality reduction
- Revelation of data structure (*clustering*, *correlations*)
- Interpretation of differences among samples

Further processing

- Quantitative Analysis
- Discriminant Analysis

Bibliography

Vandeginste, Massart,Buydens, De Jong, Lewi, Smeyers-Verbeke

Handbook of Chemometric and Qualimetric

Chapters 17, 31

Elsevier Science BV, Amsterdam, 1998

M. J. Adams

Chemometric in Analytical Spectroscopy

Chapter 3

Royal Society of Chemistry, Cambridge, 1995

J. E. Jackson

A User's Guide to Principal Components

John Wiley & Sons Inc. Hoboken, New jersey, 2003