

Faculty: BioScienze e Tecnologie Agro-Alimentari e Ambientali
MASTER DEGREE IN FOOD SCIENCE AND TECHNOLOGY
I YEAR

Course:

**EXPERIMENTAL DESIGN AND
CHEMOMETRICS IN FOOD**
(5 credits – 38 hours)

Teacher: Marcello Mascini
(mmascini@unite.it)

The Teacher is available to answer questions at the end of the lesson, or on request by mail

The course is split in 4 units

UNIT 1: Univariate analysis

Data, information, models, data types, analytical representation of data

Calibration and regression, Introduction to Statistics

Average & Variance

The Normal distribution, theory of measurement errors, the central limit theorem and the theorem of Gauss

Maximum likelihood, method of least squares, Generalization of the method of least squares

Polynomial regression, non-linear regression, the χ^2 method, Validation of the model

UNIT 2: Multivariate analysis

Correlation

Multiple linear regression

Principal component analysis (PCA)

Principal component regression (PCR) and Partial least squares regression - (PLS)

UNIT 3: Design of Experiments

Basic design of experiments and analysis of the resulting data

Analysis of variance, blocking and nuisance variables

Factorial designs

Fractional factorial designs

Overview of other types of experimental designs (Plackett–Burman designs, D-optimal designs, Supersaturated designs, Asymmetrical designs)

Response surface methods and designs

Applications of designed experiments from various fields of food science

UNIT 4: Elements of Pattern recognition

cluster analysis

Normalization

The space representation (PCA) Examples of PCA

Discriminant analysis (DA) PLS-DA

Examples of PLS-DA

UNIT 2: Multivariate Analysis

Correlation

Multiple linear regression

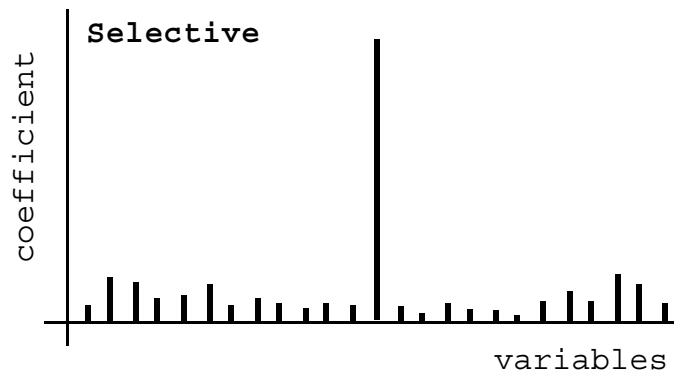
Principal component analysis (PCA)

Principal component regression (PCR) and

Partial least squares regression - (PLS)

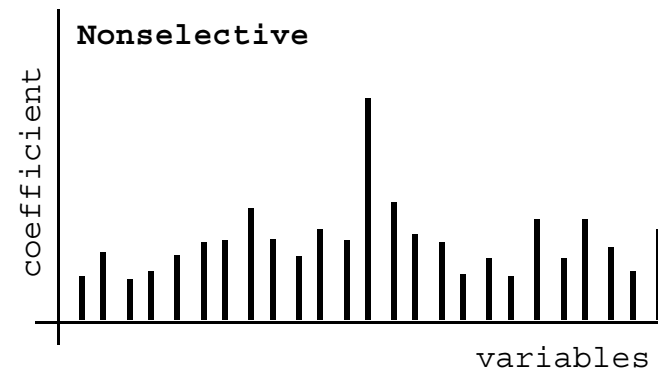
selective and non-selective measurements

- The measurements can be selective or non-selective
 - Selective: the observation is driven by one variable
 - Non Selective: The observation is driven by many variables
- The non selective measurements are the objects of the multivariate analysis



↓

$$z \cong k_j \cdot C_j$$



↓

$$z = \sum_i k_i \cdot C_i$$

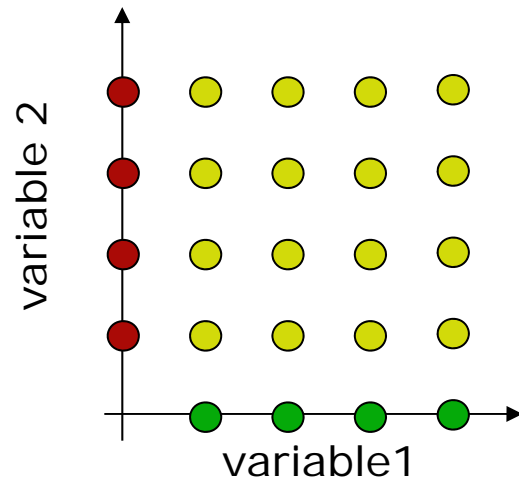
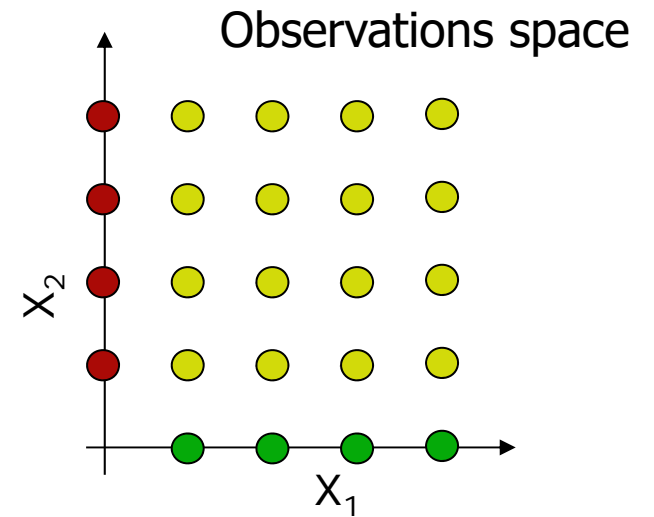
Non selective measurements

- Example:
 - Spectroscopy
 - At a given frequency the absorbance is influenced by more than one molecule
 - Gas chromatography
 - Compounds with similar elution time can contribute to chromatographic peak
 - Sensors
 - The sensor response is given by the combination of different compounds that interfere with the sensors depending on concentration and affinity

Variables and observations space : selective measurements

$$Y_1 = aX_1 + bX_2$$

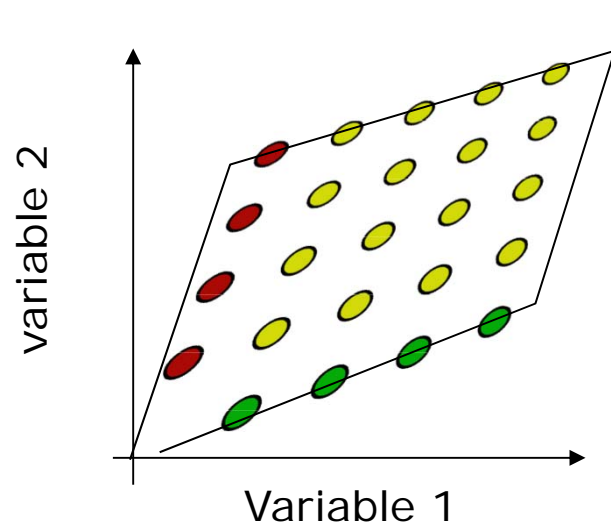
$$Y_2 = cX_1 + dX_2$$



$$K = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}$$

Correlation = 1 - det(K) = 0

Variables and observations space : selective measurements

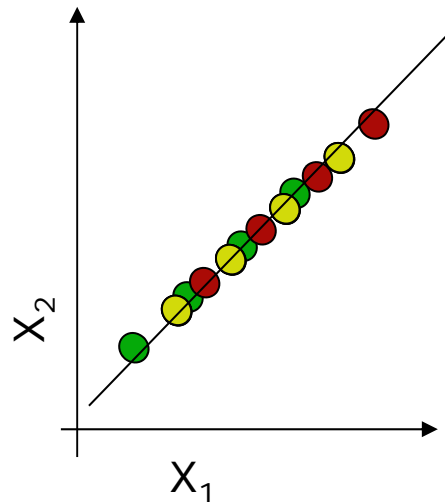


Partial correlation

$$0 < c < 1$$

$$K = \begin{vmatrix} a & b \\ c & d \end{vmatrix}$$

$$a, b, c, d \neq 0$$



Total correlation

$$c = 1$$

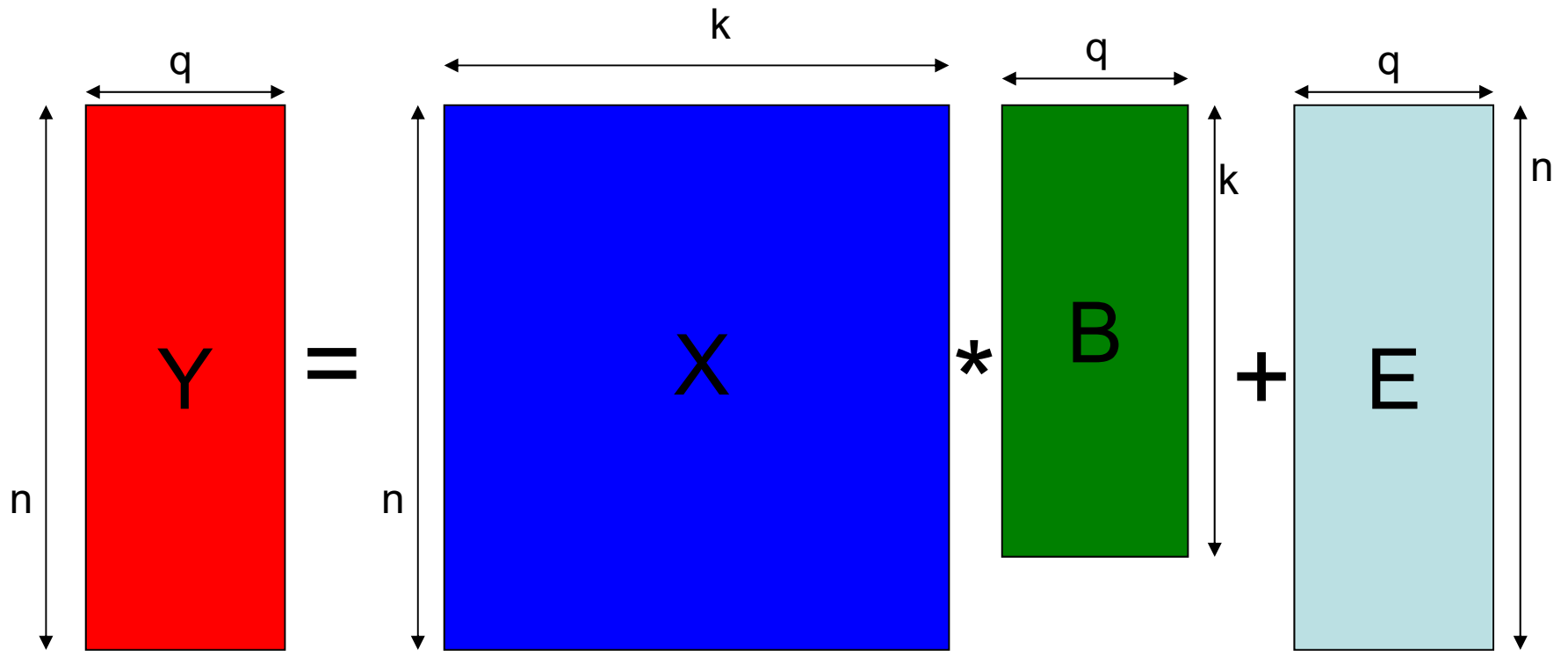
$$K = \begin{vmatrix} a & b \\ c & d \end{vmatrix}$$

$$a \cdot d - b \cdot c = 0$$

Multiple Linear Regression

- Multiple linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable from two or more independent variables. The independent variables can be continuous or categorical .
- Multiple linear regression analysis makes several key assumptions:
- Linear relationship.
- Multivariate normality.
- No or little multicollinearity.
- No auto-correlation.
- Homoscedasticity (The variance around the regression line is the same for all values of the predictor variable (X)).

Multiple Linear Regression



$k = n^\circ$ observations
 $n = n^\circ$ measurements
 $q = n^\circ$ variables

$$Y = XB + E$$

Multiple Linear Regression

- as for the univariate event we use two steps :
 - Calibration: using known Y and X we determine the matrix B (the slope) B
 - Procedure: known the matrix B we can have an optimized estimation of X by measuring Y
- Calibration:
 - Known X and Y the best estimation of B is given by the Gauss-Markov theorem :

$$B_{MLR} = X^+ \cdot Y$$

- If the matrix X has the maximum rank we can calculate the pseudoinverse in this way :

$$B_{MLR} = \left(X^T \cdot X \right)^{-1} \cdot X^T \cdot Y$$

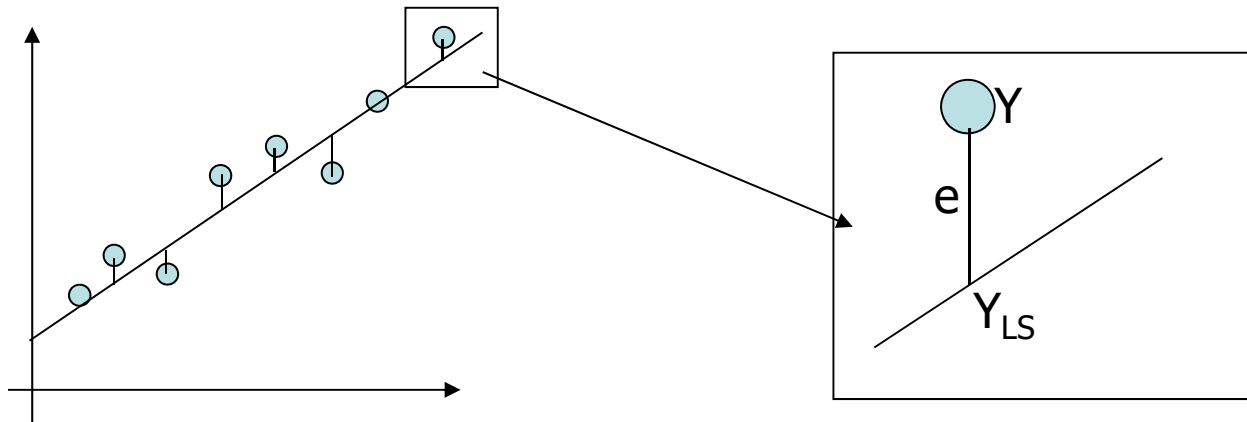
- **It means that every observation is independent from each other**

MLR meaning

In a linear regression model in which the errors have expectation zero and are uncorrelated and have equal variances, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator.

- In practice B_{MLR} maximize the correlation between X and Y
- Geometrically the Y orthogonal projection In a subspace of X
- Ω is a matrix in a subspace of X

$$Y_{MLR} = X \cdot B_{MLR} = X \cdot (X^T \cdot X)^{-1} \cdot X^T \cdot Y = \Omega \cdot Y$$



MLR Limitations

Regression analysis is concerned with developing the linear regression equation by which the value of a dependent variable Y can be estimated given a value of an independent variable X . If simple regression analysis is used, the assumptions for this technique should be satisfied. The assumption required to develop the linear regression equation and to estimate the value of dependent variable by point estimation is: 1. The relationship between the two variables is linear. 2. The value of the independent variable is a set at various values, while the dependent variable is a random variable. 3. The conditional distributions of the dependent variable have equal variances.

If any interval estimation or hypothesis testing is done, additional required assumptions are: 1. **Successive observations of the dependent variable are uncorrelated.**

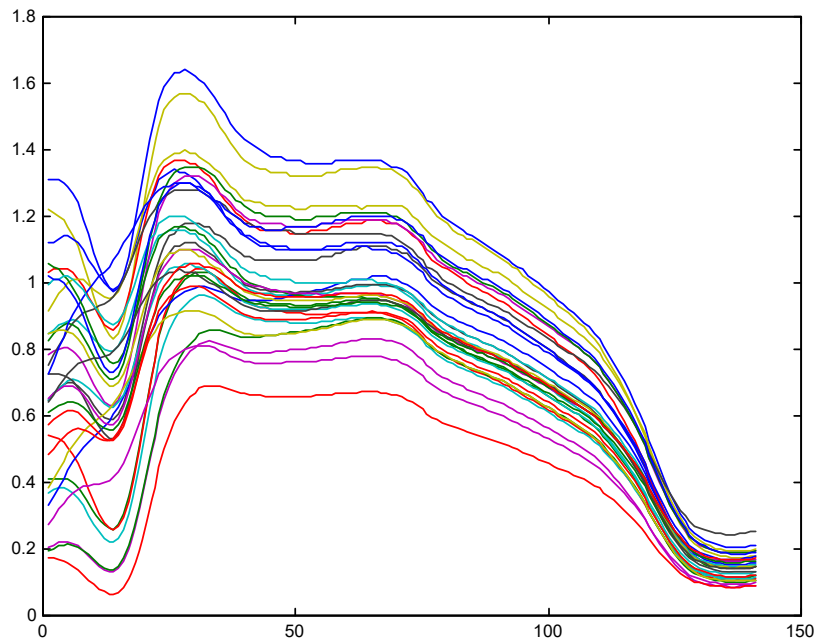
2. The conditional distributions of the dependent variable are normal distributions.

- **IF the observations of the dependent variable are correlated we have to find a method to transform them in uncorrelated observations**

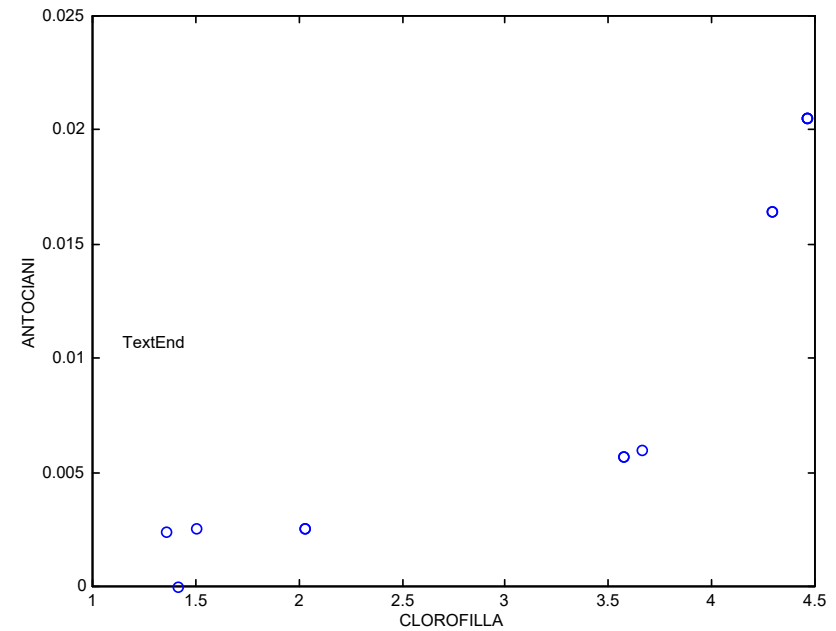
Example

Chlorophyll and anthocyanins in peaches using Vis-NIR

- `sptectra(Y)`

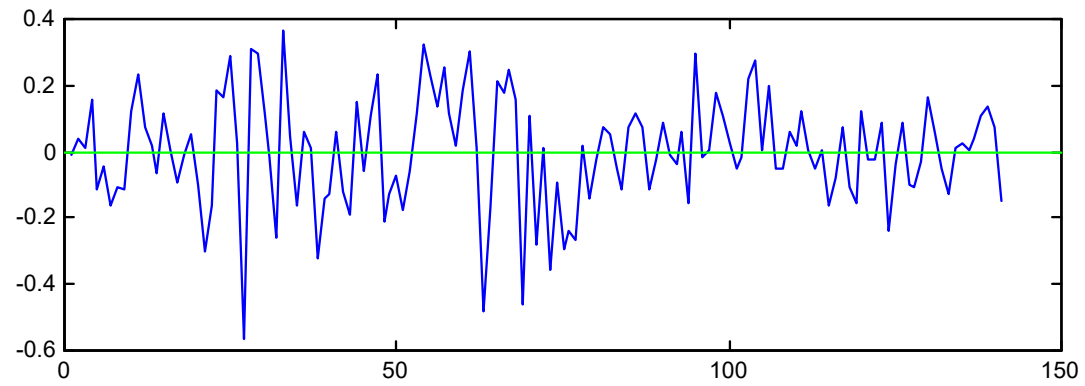
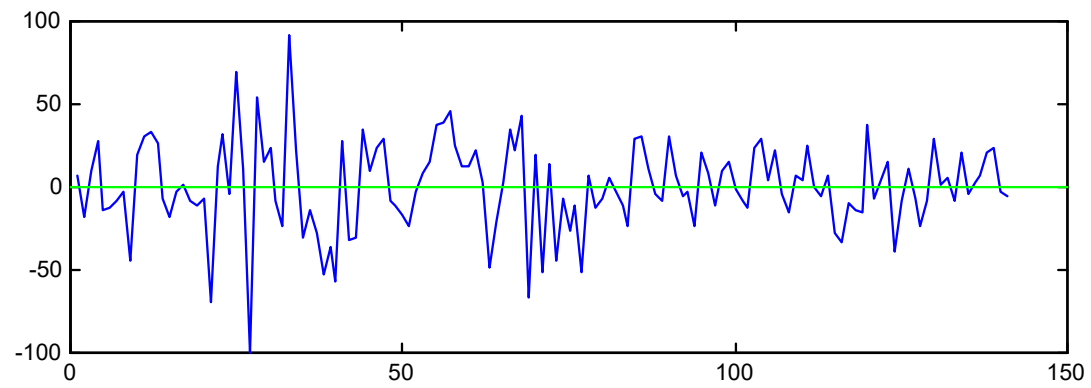


- Chlorophyll and anthocyanins (**X**)



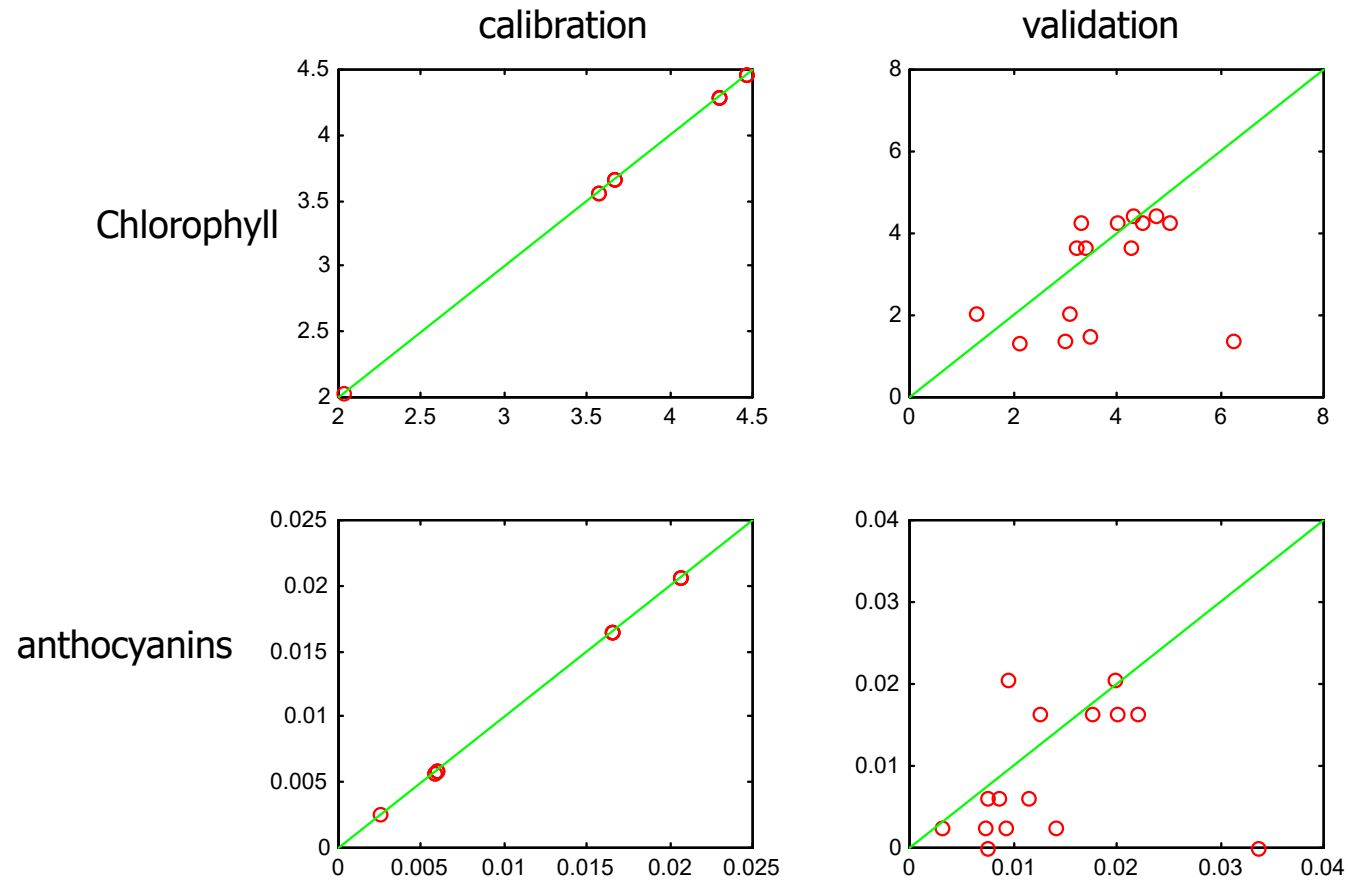
results

- Matrix coefficient **B**



results

- Y_{LS} and Y comparison
 - Scatter plot: x Axis: true value; y Axis : estimated value



Principal components analysis (PCA)

Analysis of Variance

PCA and diagonalization of the covariance matrix

Scores and loadings

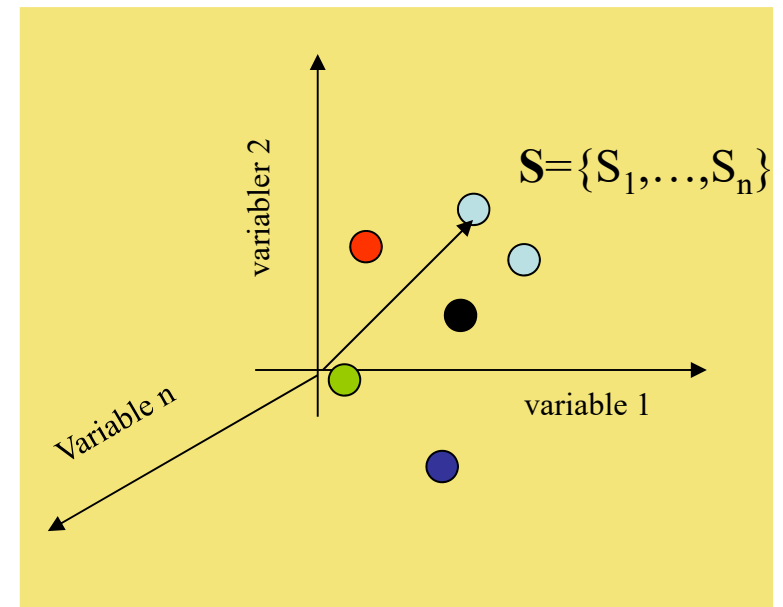
residual matrix

Applications to image analysis

Applying the multivariate regression: Principal Components Regression (PCR)

Observations space

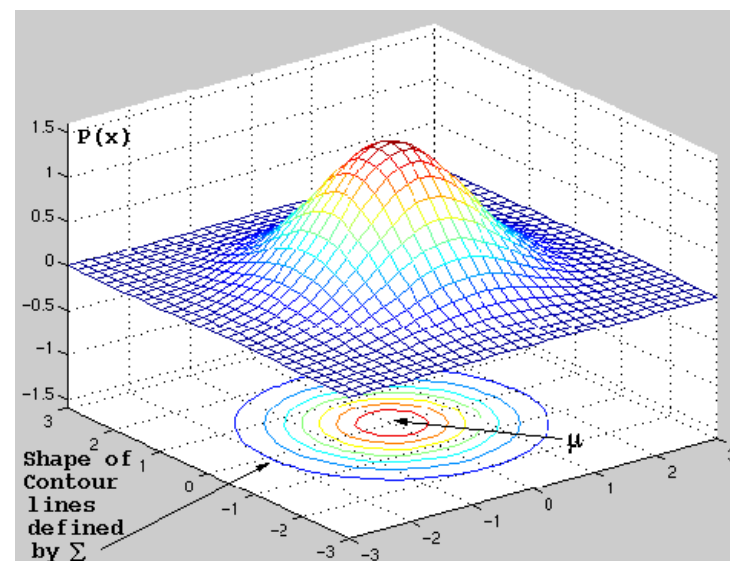
- Each multivariate measurement is represented by a vector in a space to N dimensions
- N is equal to the size of the vector that expresses the observation
- The statistical distribution of points (vectors) defines the properties of the entire data set.
- For each multivariate data we can define a PDF multivariate.
- Important: observations that describe similar samples are represented by closest points then mutual relation between distance and similarity between samples (Hypothesis of pattern recognition)



Multivariate statistics

- the fundamental descriptors for Univariate distribution :
 - Average scalar \Rightarrow vector
 - Variance scalar \Rightarrow matrix (covariance matrix)
 -
- The normal distribution defined in univariate approach it keeps its importance in multivariate approach

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \dots & \sigma_{1d} \\ \vdots & \dots & \vdots \\ \sigma_{d1} & \dots & \sigma_{dd} \end{bmatrix}$$
$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi} \sqrt{|\Sigma|}} \exp \left[-\frac{1}{2} (\mathbf{x} - \mu)^t \Sigma^{-1} (\mathbf{x} - \mu) \right]$$



Covariance matrix

- In probability theory and statistics, a covariance matrix (also known as dispersion matrix or variance–covariance matrix) is a matrix whose element in the i, j position is the covariance between the i th and j th elements of a random vector. A random vector is a random variable with multiple dimensions. Each element of the vector is a scalar random variable. Each element has either a finite number of observed empirical values or a finite or infinite number of potential values. The potential values are specified by a theoretical joint probability distribution. Because the covariance of the i th random variable with itself is simply that random variable's variance, each element on the principal diagonal of the covariance matrix is just the variance of each of the elements in the vector. Every covariance matrix is symmetric. In addition, every covariance matrix is positive semi-definite.
- The covariance matrix can be done by : $\text{COV}(xy) = x^T y$

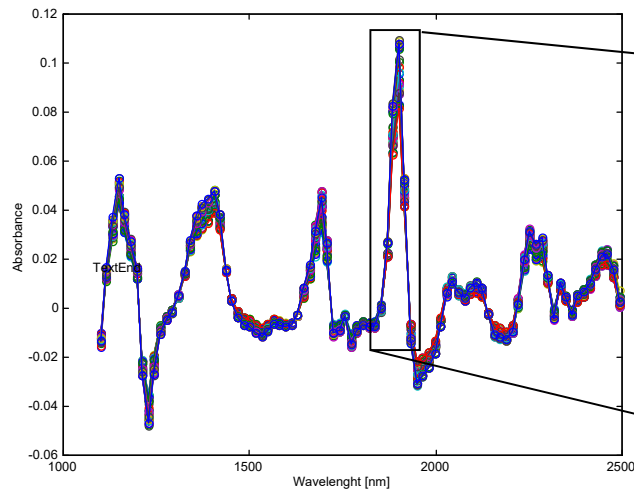
Multicollinearity

- Multicollinearity (also co-linearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. In this situation the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors. That is, a multiple regression model with correlated predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others.
- In case of perfect multicollinearity the design matrix X has less than full rank, and therefore the moment matrix $X^T X$ cannot be inverted. Under these circumstances, for a general linear model $Y = cX + E_r$, the ordinary least-squares estimator does not exist.

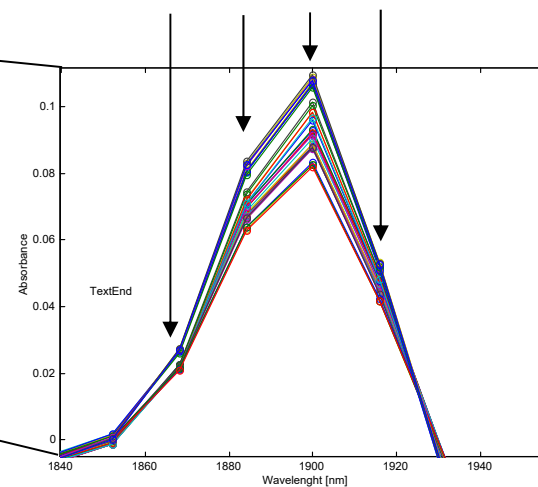
Co-linearity example

- In an optical spectrum the spectral lines cover a range of wavelengths, this interval is generally covered by more spectral channels, so that more variables combine to form a spectral line.
- If the line is proportional to a characteristic of the sample (eg. Glucose concentration) all the spectral channels related to the line will be proportional to the sample characteristic, and then the relative variables (columns in the data matrix) will become collinear.
- co-linear variables depend quantitatively by the sample characteristics

NIR of fruits



Collinear variables

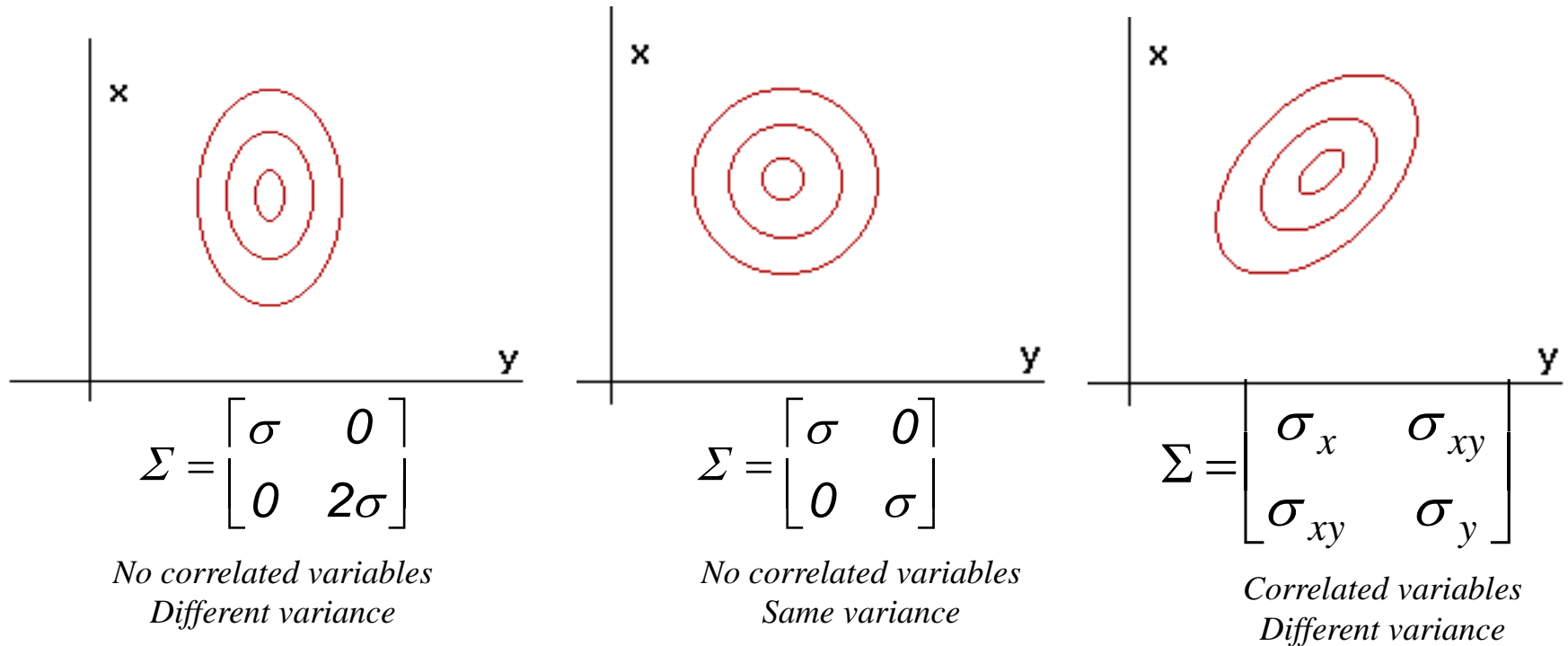


Covariance matrix and co-linearity

- The co-linearity is expressed by the covariance matrix.
- In case of co-linearity the non-diagonal terms of the covariance matrix are nonzero.
- Remove the co-linearity it means manipulating the covariance matrix in diagonal form by introducing new latent variables.
- The principal component analysis technique allows, among other things, to obtain this result!!

Example of covariance matrix and points probability

- Example of bivariate distribution



Multivariate PDF and covariance matrix

- The multivariate distribution only makes sense if the covariance matrix describes the parameters correlated with each other, that is, if the matrix is not diagonal.
- In fact, for two quantities (x and y) unrelated and independent the probability to observe simultaneously the value of x and y is simply the product of the two univariate distributions:

$$P(x, y) = P(x) \cdot P(y)$$

The covariance matrix in canonical form

- The covariance matrix can be written in diagonal form with an appropriate change of the reference system.
- Such a reference system corresponds to the eigenvectors of the covariance matrix, ie the main ellipse constructed as quadratic form from the covariance matrix itself.
- This operation makes variables uncorrelated and the PDF as a product of the univariate PDF .
- On the other hand the new variables are no longer physical observables (object of measurement) but are linear combinations of these.
- The new variables are called Principal Components and the set of calculation procedures and interpretation of the main components is called principal component analysis (PCA)

$$a \cdot x^2 + 2b \cdot xy + c \cdot y^2 = \begin{bmatrix} x & y \end{bmatrix} \cdot \begin{bmatrix} a & b \\ b & c \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\Rightarrow \lambda_1 \cdot u^2 + \lambda_2 \cdot w^2 = \begin{bmatrix} u & w \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \cdot \begin{bmatrix} u \\ w \end{bmatrix}$$

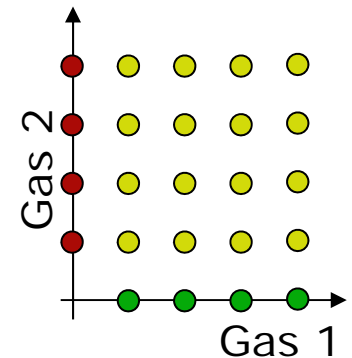
Dimension of the data set

- If the variables of a multivariate phenomena have a certain degree of correlation then the representative vectors of the phenomenon will occupy only a portion of the observation space .
- So a variable of size N will lie in a space of smaller dimension

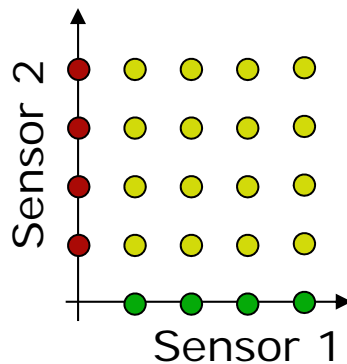
Example: linear sensors

$$\begin{cases} s_1 = k_{11} \cdot g_1 + k_{12} \cdot g_2 \\ s_2 = k_{21} \cdot g_1 + k_{22} \cdot g_2 \end{cases}$$

Independent variables

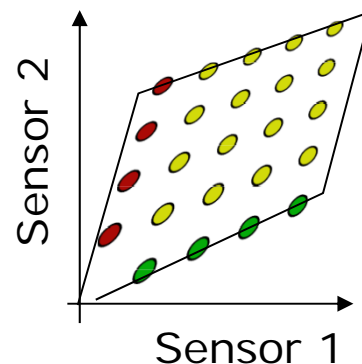


Specific sensors
 $k_{12}=k_{21}=0$



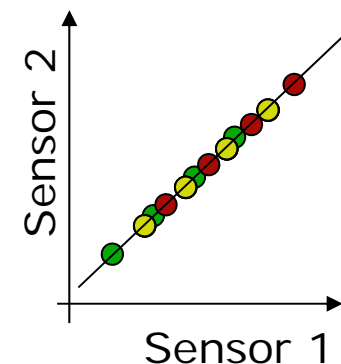
$C=0$ Dim=2

No specific sensors but
 $k_{11}; k_{12}; k_{21}; k_{22}$ different



$C > 0$ and < 1 Dim

No specific but equal k

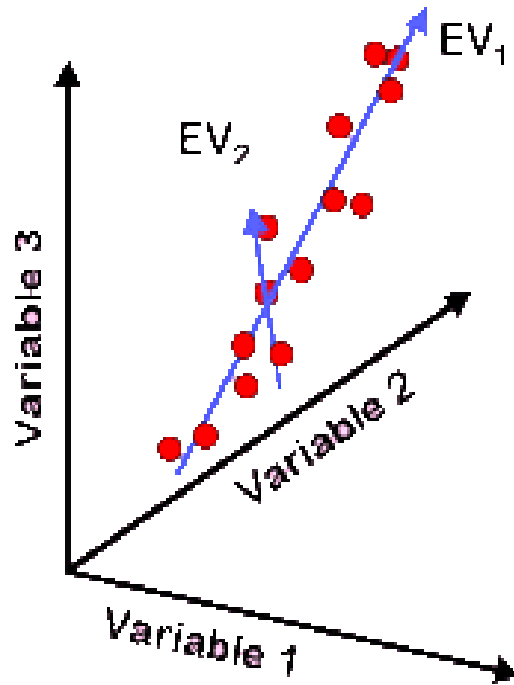


$C=1$ Dim=1

Principal Component Analysis

- The purpose of the PCA is the representation of a data set having covariance matrix not diagonal and with a space of smaller dimension in which the same data are represented by a diagonal covariance matrix.
- The diagonalization is achieved with a coordinate rotation in the base of the eigenvectors (principal components)
- For each eigenvector it is associated an eigenvalue which corresponds to the variance of the associated component. If the original variables were partially correlated some eigenvalues have a negligible value.
- In practice the corresponding eigenvectors can be ignored by limiting the representation only to eigenvectors with the largest eigenvalues.
- Since the covariance matrix in the base of the main components is diagonal, the total variance is the sum of the variances of the individual components.

PCA procedure



- Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

PCA

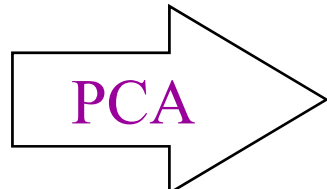
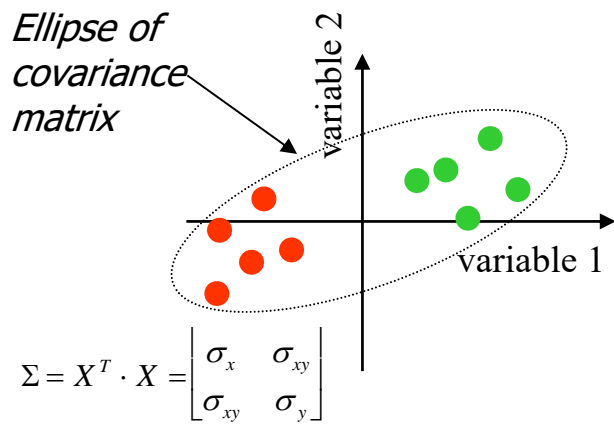
- PCA is mostly used as a tool in exploratory data analysis and for making predictive models. PCA can be done by eigenvalue decomposition of a data covariance (or correlation) matrix or singular value decomposition of a data matrix, usually after mean centering (and normalizing or using Z-scores) the data matrix for each attribute. The results of a PCA are usually discussed in terms of component scores, sometimes called factor scores (the transformed variable values corresponding to a particular data point), and loadings (the weight by which each standardized original variable should be multiplied to get the component score).

PCA

- PCA is the simplest of the true eigenvector-based multivariate analyses. Often, its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance in the data. If a multivariate dataset is visualized as a set of coordinates in a high-dimensional data space (1 axis per variable), PCA can supply the user with a lower-dimensional picture, a projection or "shadow" of this object when viewed from its (in some sense; see below) most informative viewpoint. This is done by using only the first few principal components so that the dimensionality of the transformed data is reduced.
- PCA is closely related to factor analysis. Factor analysis typically incorporates more domain specific assumptions about the underlying structure and solves eigenvectors of a slightly different matrix.
- PCA is also related to canonical correlation analysis (CCA). CCA defines coordinate systems that optimally describe the cross-covariance between two datasets while PCA defines a new orthogonal coordinate system that optimally describes variance in a single dataset.

PCA

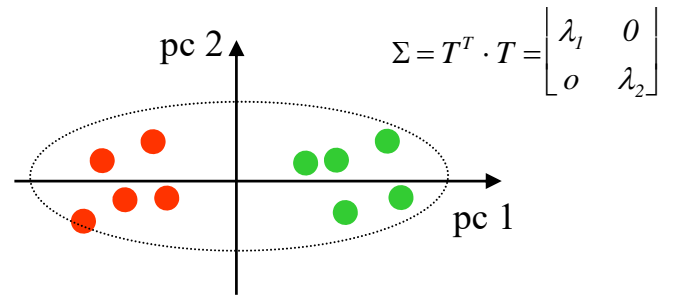
Observation space



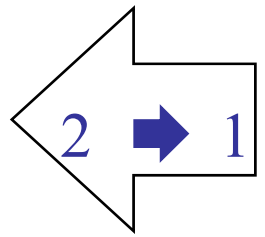
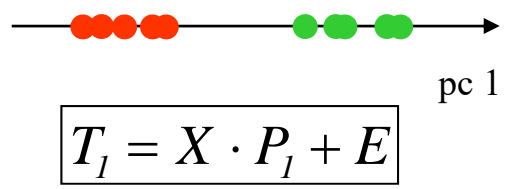
$$\Sigma = X^T \cdot X \Rightarrow \Lambda \cdot P^T$$

$$T = X \cdot P ; X = T \cdot P^T$$

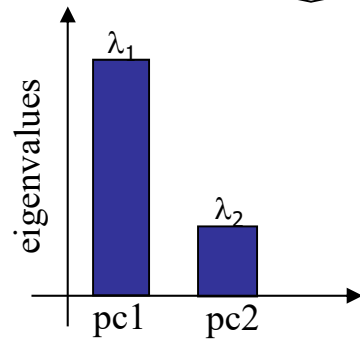
PCA space



Reduced space



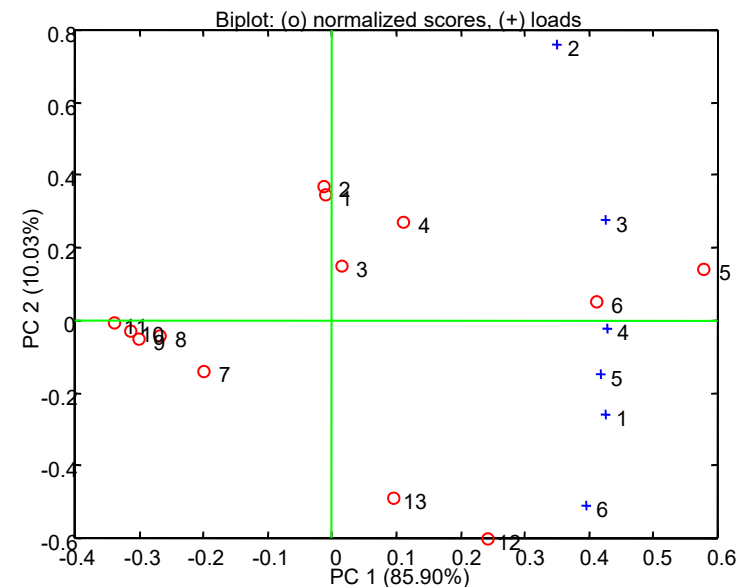
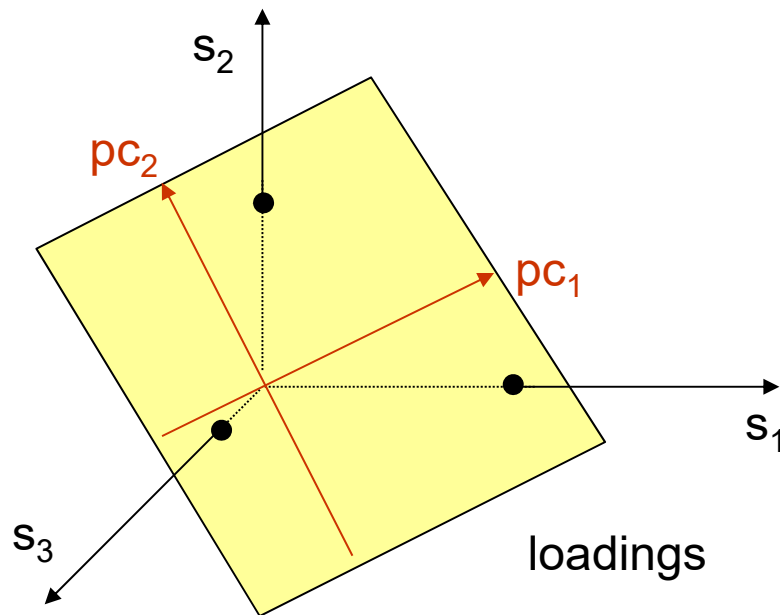
Dimensions reduction



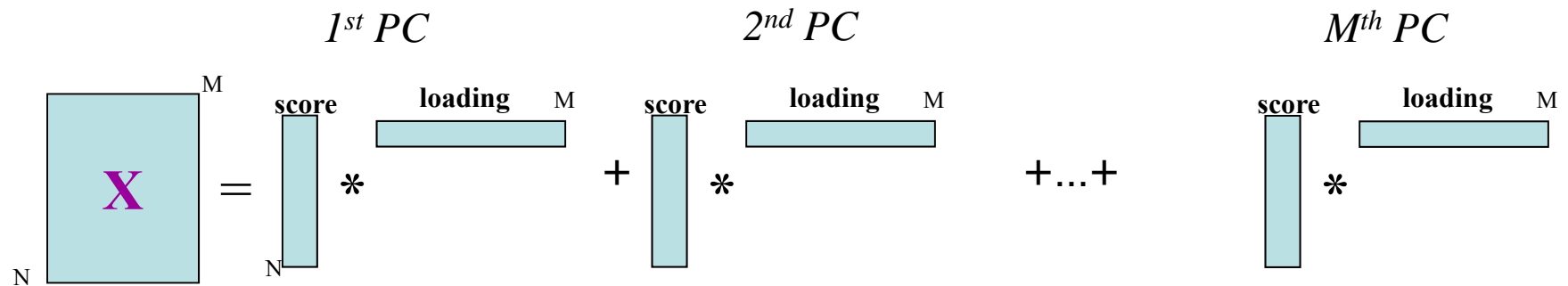
eigenvalues: a PC has a greater information content than an other

PCA: scores e loadings

- The new coordinates of the vectors corresponding to the observations (the rows of the matrix x) in the base of the principal components are called scores
- The coefficients of the linear combinations that define the principal components are called loadings
- The loading therefore provides a measure of the contribution of each observable to the principal components
- The loadings are also represented as scores as they are the projection of the original axes in the subspace identified the principal components, and scores and loadings can be plotted together



PCA matrix Decomposition

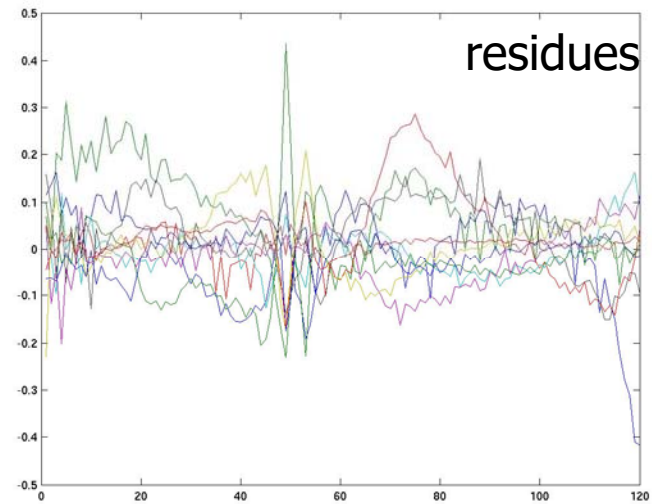
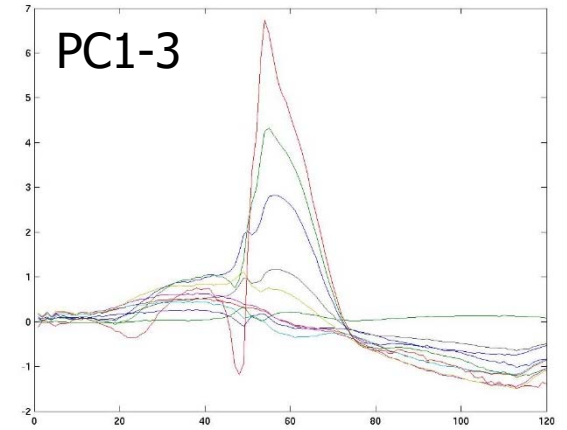
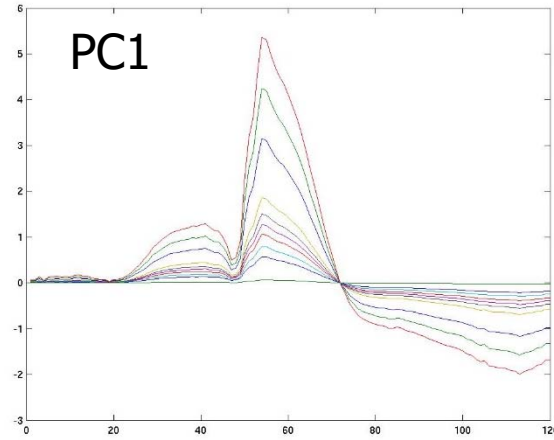
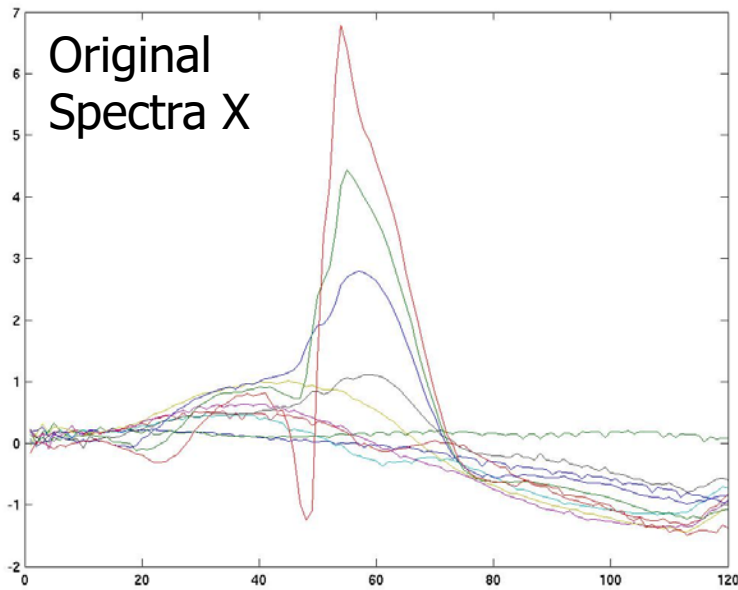


$$X_{nm} = S_{np} \cdot L_{pm}^T + \text{Residual}$$

PCA, correlation and noise

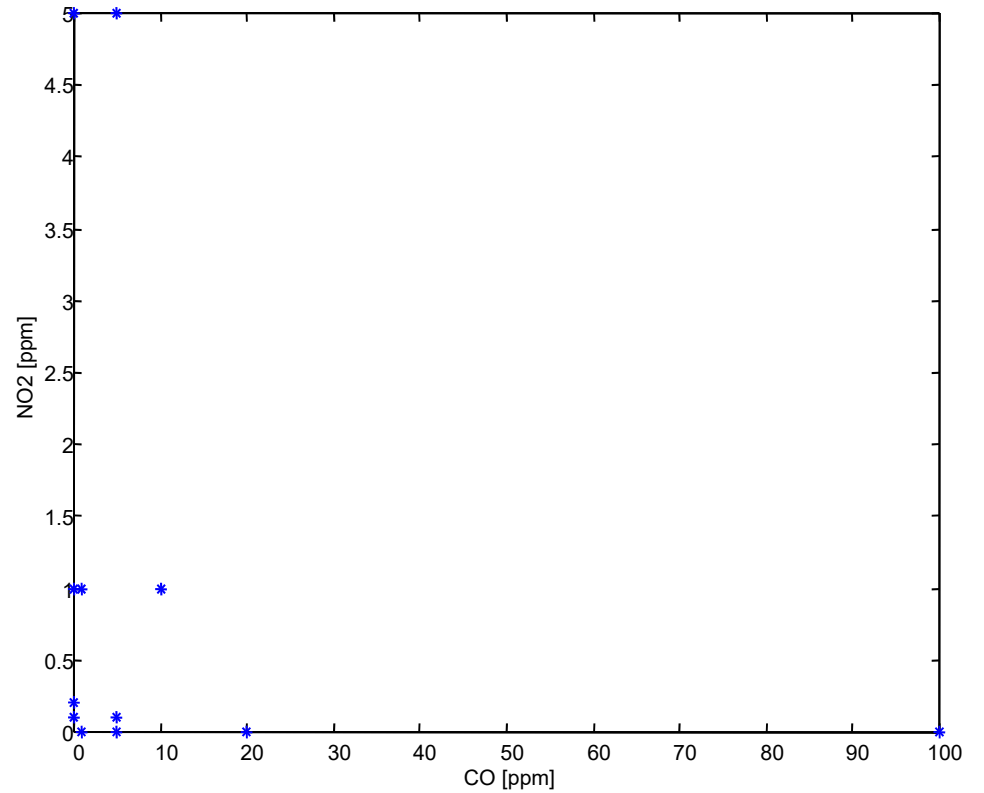
- Noise is an additional stochastic term that belongs to every observation.
- The noise is the term that makes the measurement a statistical operation.
- The principal components describe the directions of maximum correlation between the data, for which the higher-order PC are oriented towards the directions of maximum correlation and those of lower order towards the poor correlation directions
- Decomposing the major components of higher order means holding the maximum correlation directions and remove those that are no-correlated. In no-correlated directions where there is the noise
- The PCA therefore is a method for reducing the noise in a set of multivariate data.
- example: spectroscopy, GC, ...

removing noise : Reflectance Anisotropy Spectroscopy

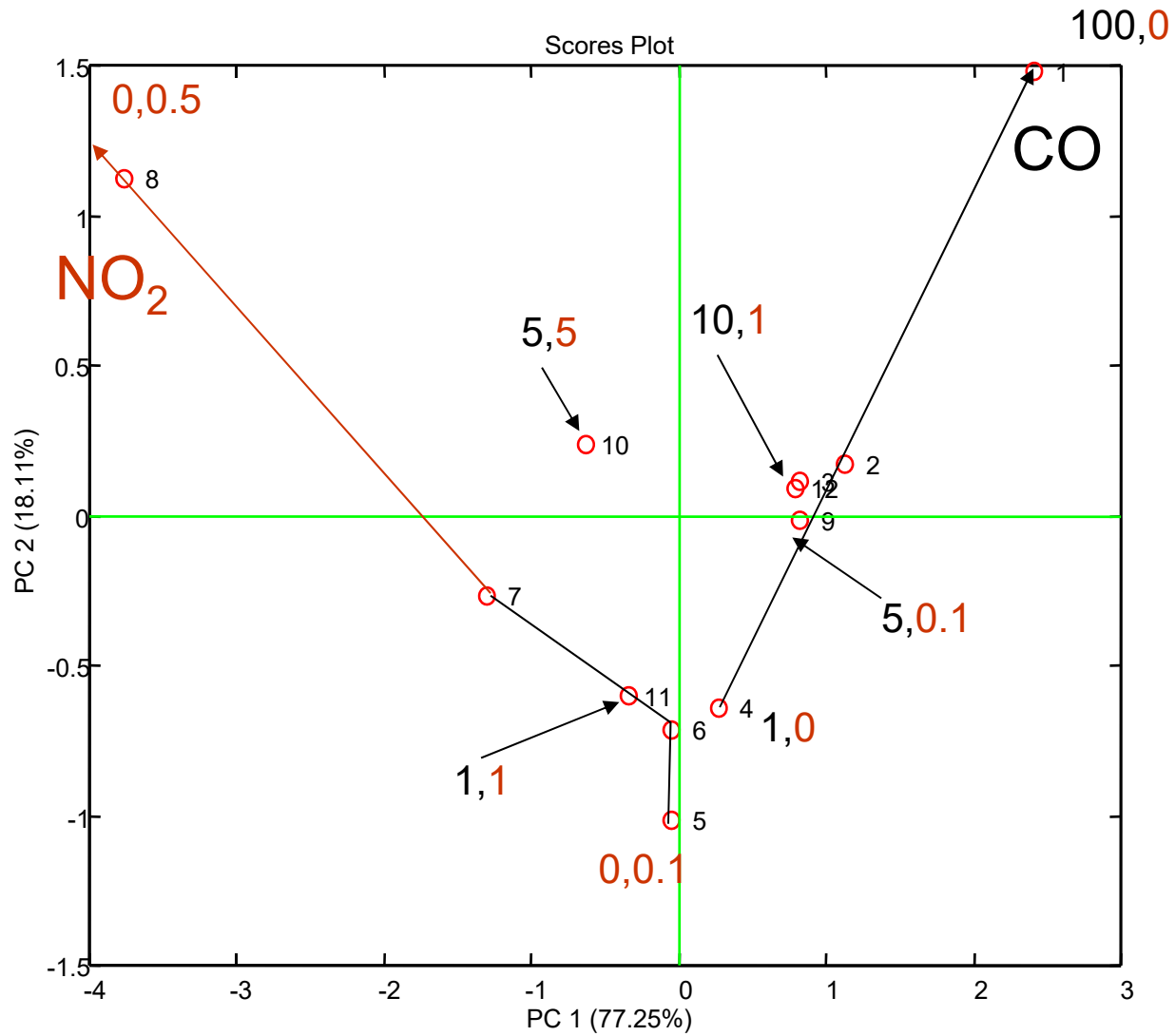


3 SnO2 sensors for 2 gas

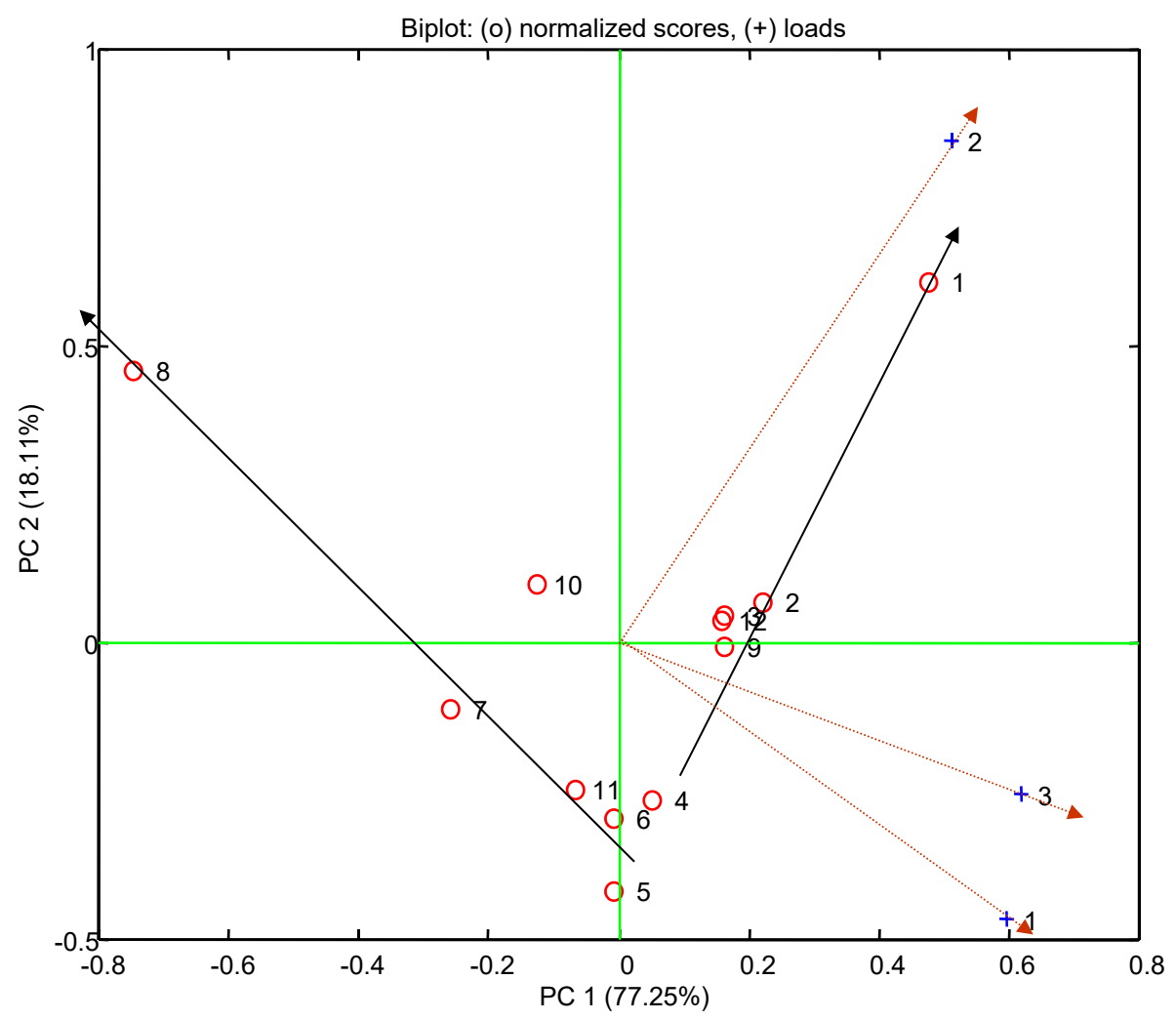
| G_r/G_i | | | CO | NO ₂ |
|-----------|------------|------------|--------|-----------------|
| 0.25482 | 0.63354 | 0.77832 | 100.00 | 0.0000 |
| 0.093899 | 0.27108 | 0.39692 | 20.000 | 0.0000 |
| 0.043410 | 0.23361 | 0.079543 | 5.0000 | 0.0000 |
| 0.0097185 | 0.043353 | -0.0021311 | 1.0000 | 0.0000 |
| -0.018016 | -0.053860 | -0.073648 | 0.0000 | 0.10000 |
| -0.028579 | 0.0023183 | -0.36593 | 0.0000 | 0.20000 |
| -0.25167 | -0.028831 | -2.4367 | 0.0000 | 1.0000 |
| -1.6960 | -0.075037 | -3.8650 | 0.0000 | 5.0000 |
| 0.057521 | 0.21072 | 0.16777 | 5.0000 | 0.10000 |
| -0.13089 | 0.13002 | -2.1376 | 5.0000 | 5.0000 |
| -0.068079 | -0.0027190 | -0.90852 | 1.0000 | 1.0000 |
| 0.050023 | 0.22771 | 0.020198 | 10.000 | 1.0000 |



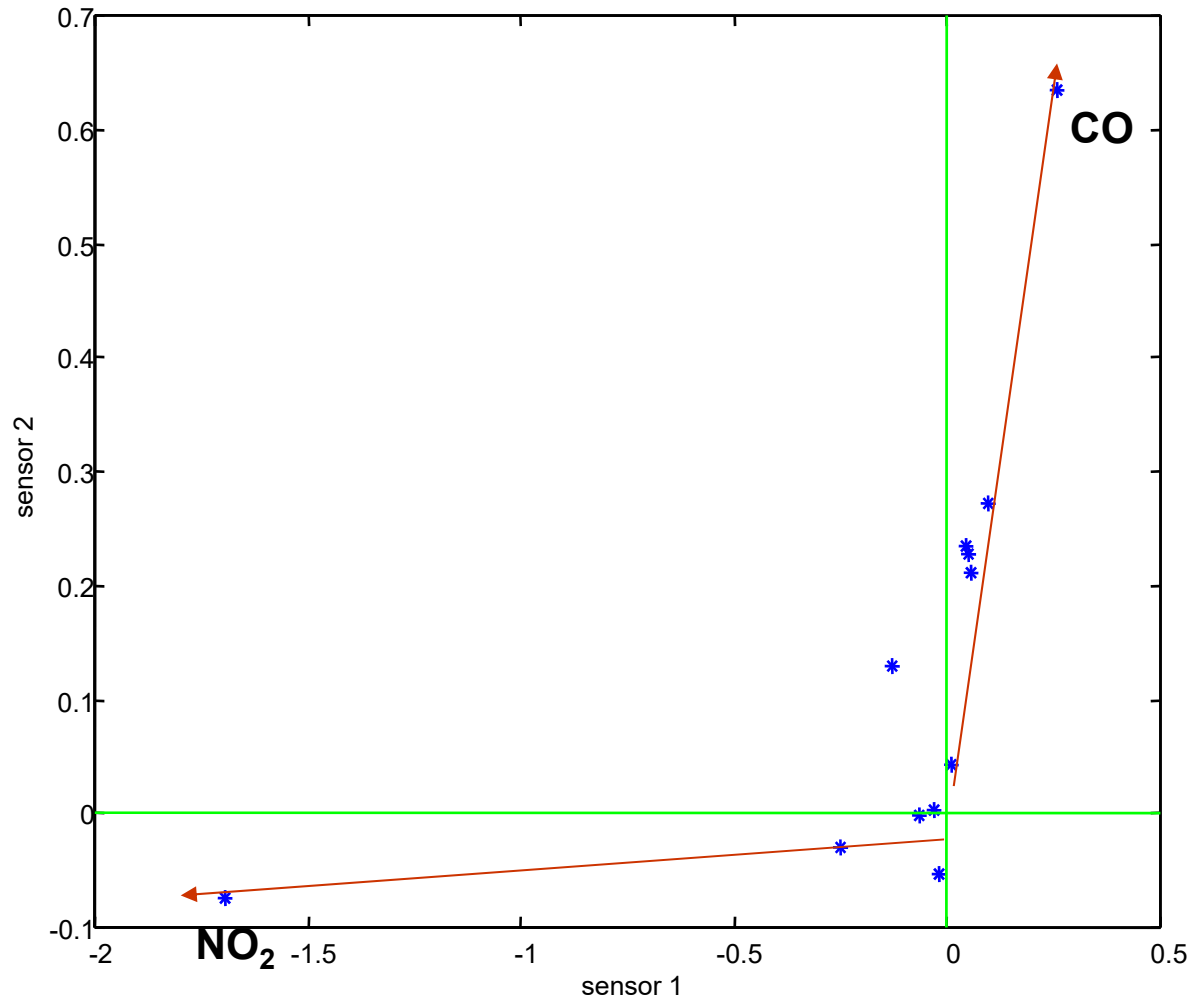
PCA score plot



PCA bi-plot



Sensor 1 vs sensor 2

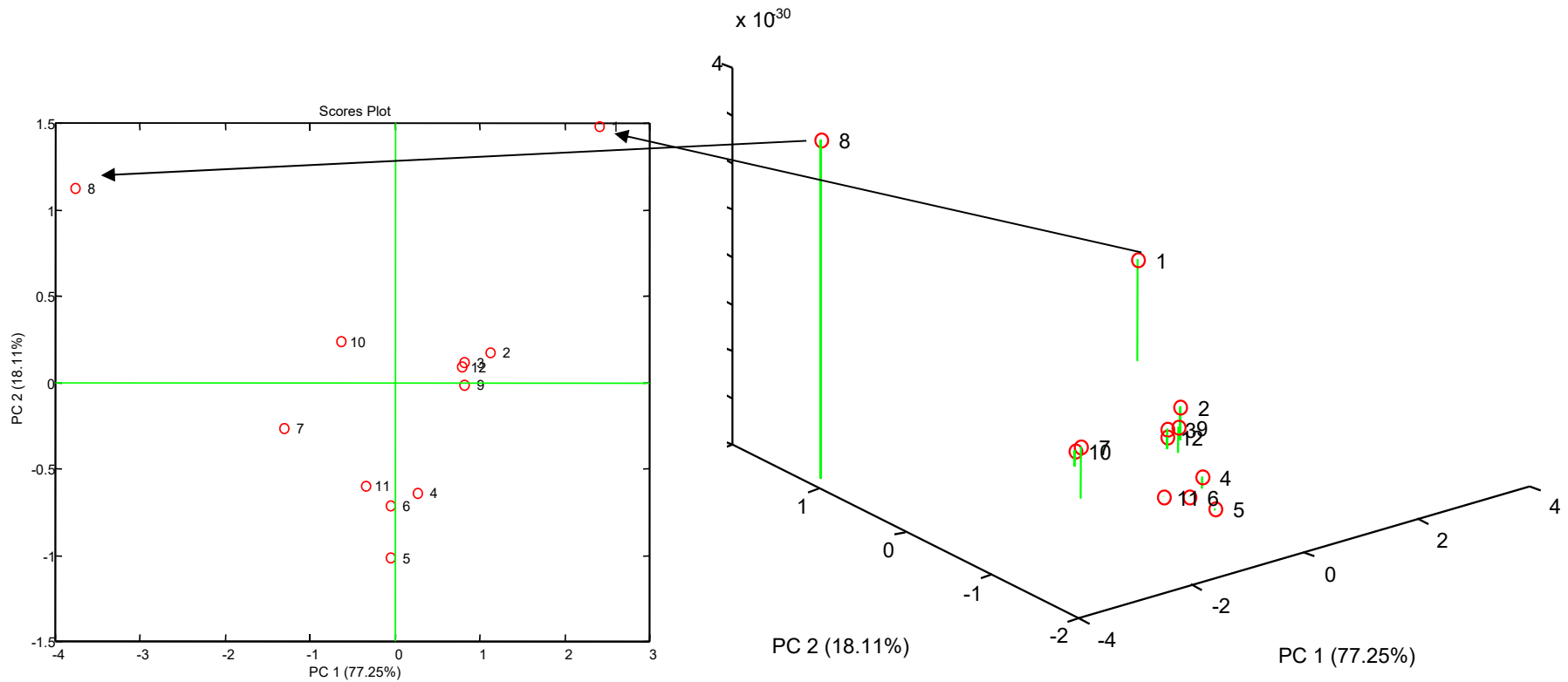


Residual of PCA representation (leverage)

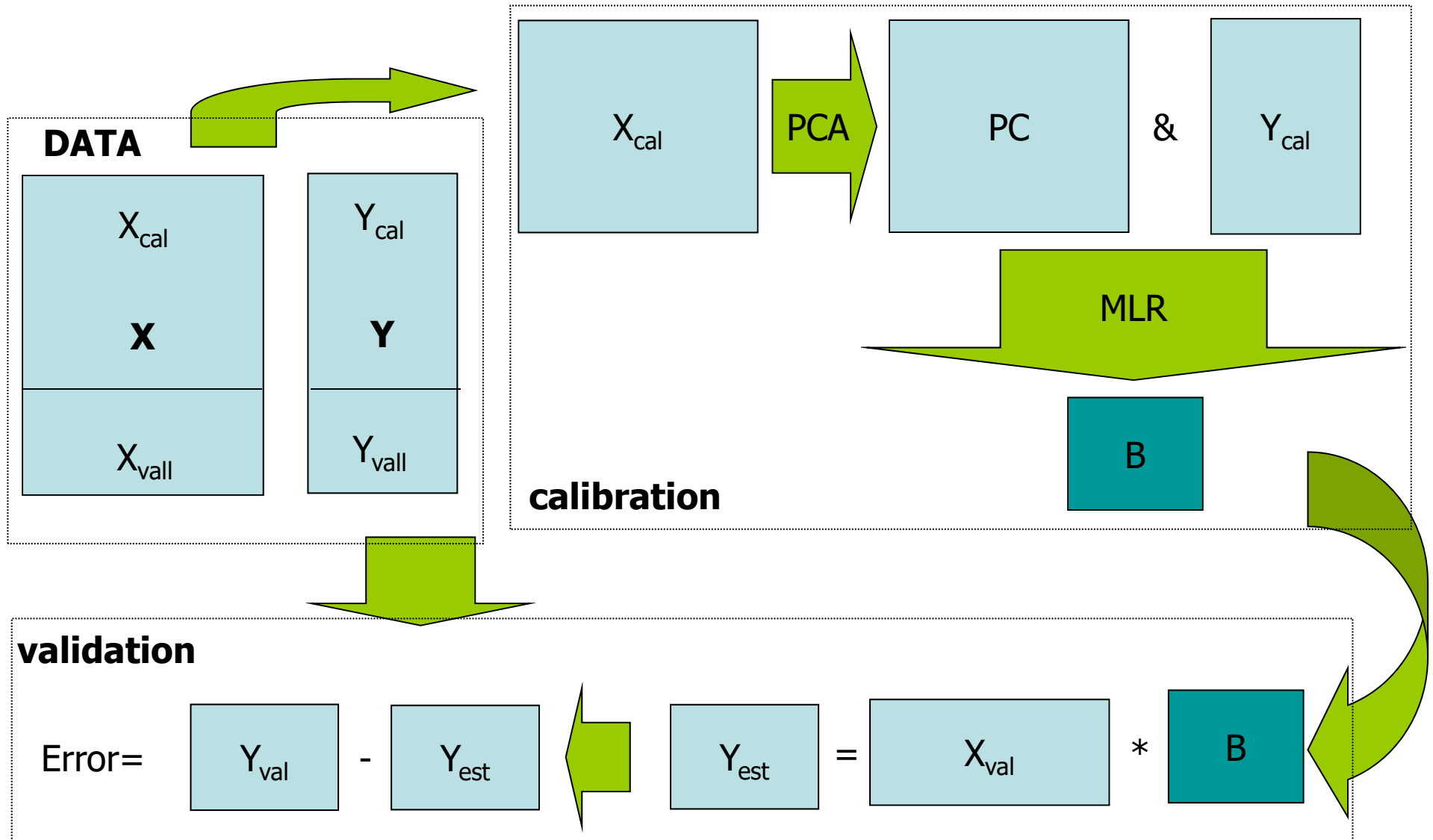
$$x_i = a \cdot s_1 + b \cdot s_2 + \dots + n \cdot s_n$$

$$x_i^{pca} = a \cdot pc_1 + b \cdot pc_2 + residual$$

Scores Plot



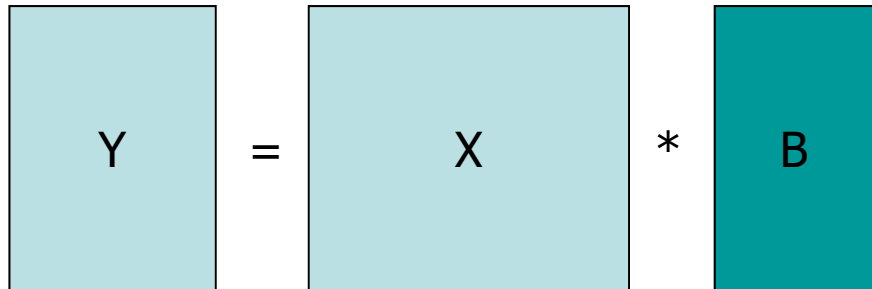
PCR procedure



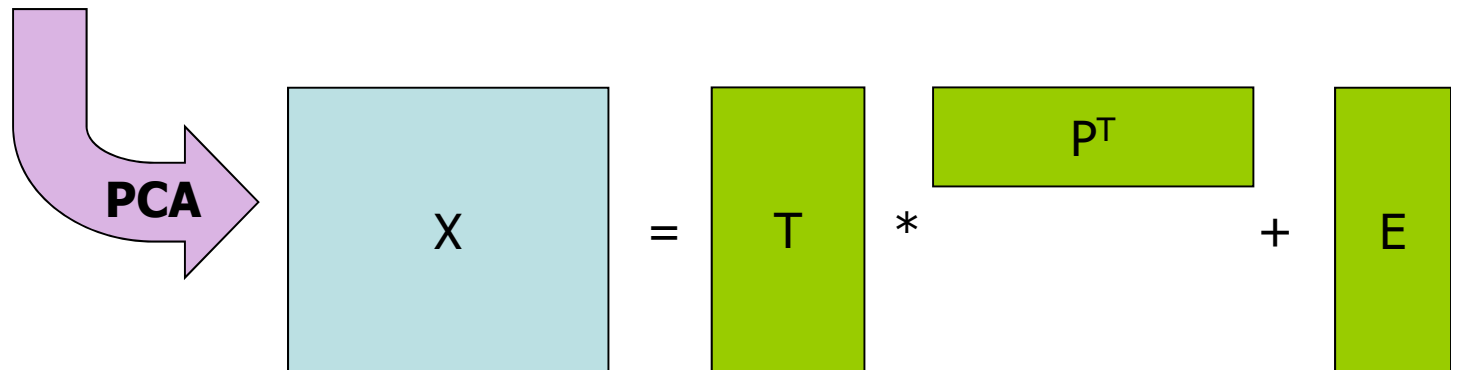
PCR algorithm

$$Y = X * B$$

Original problem

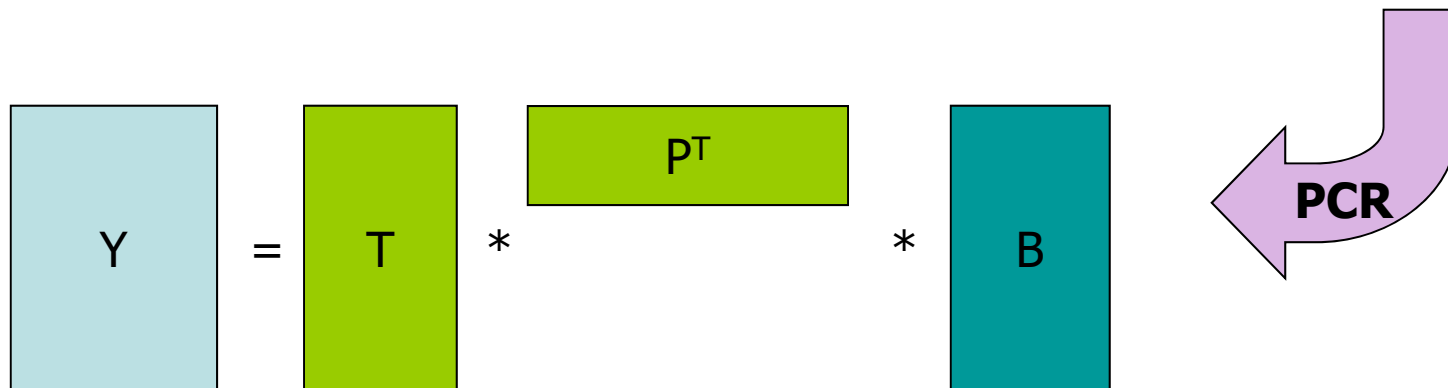


PCA

$$X = T * P^T + E$$


$$Y = T * P^T * B$$

PCR

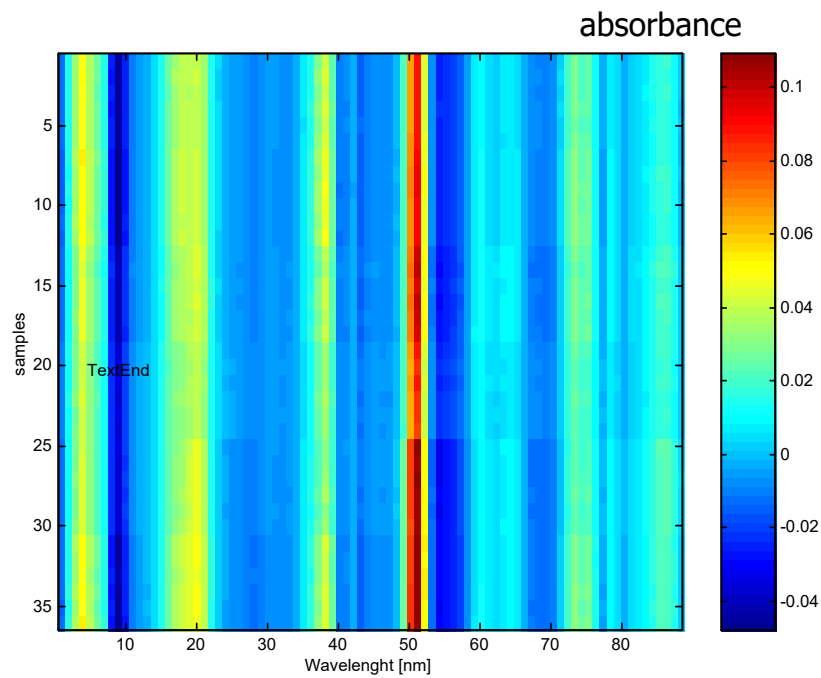


Example: NMR fruits spectra

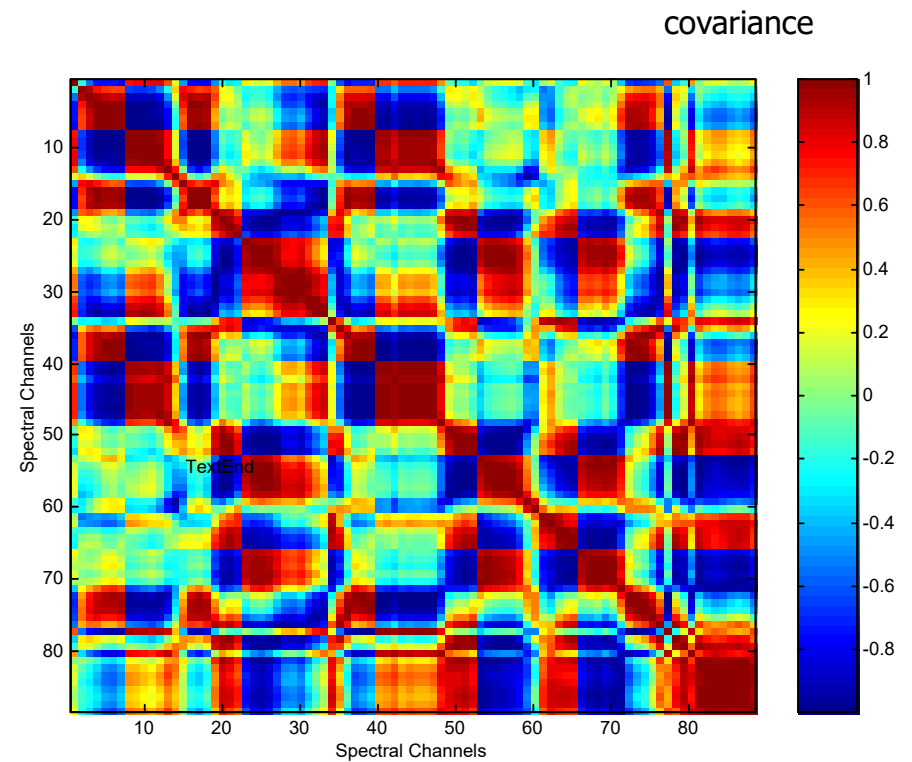
- We carried out 36 NIR spectra of fruits and we want to create a model for humidity and total acidity.
- Each spectrum is formed by 88 variables corresponding to the spectral channels in the range of 1.1-2.5 microns.
- For each fruit was measured humidity and acidity with other methods.
- We want the two parameters of Y from the spectrum X .Therefore is necessary to estimate the parameter K

$$Y_{1 \times 2} = X_{1 \times 88} \cdot K_{88 \times 2}$$

X matrix and covariance matrix

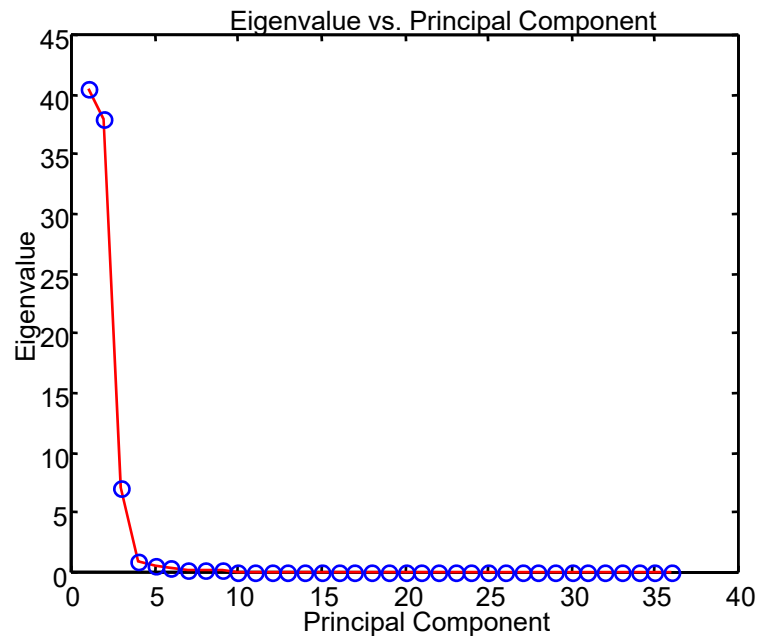


- high collinearity
- The high correlation of blocks (+ and -) correspond to the spectral lines represented by colored columns in absorbance matrix



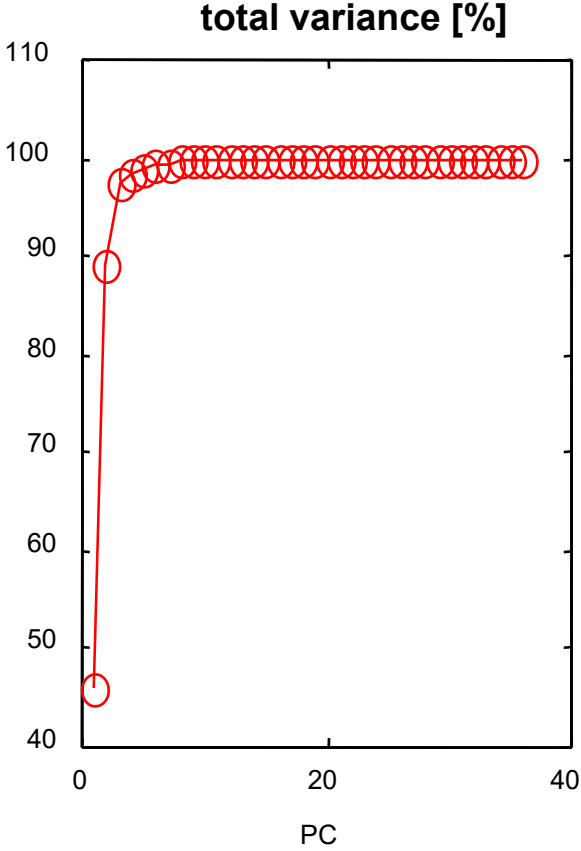
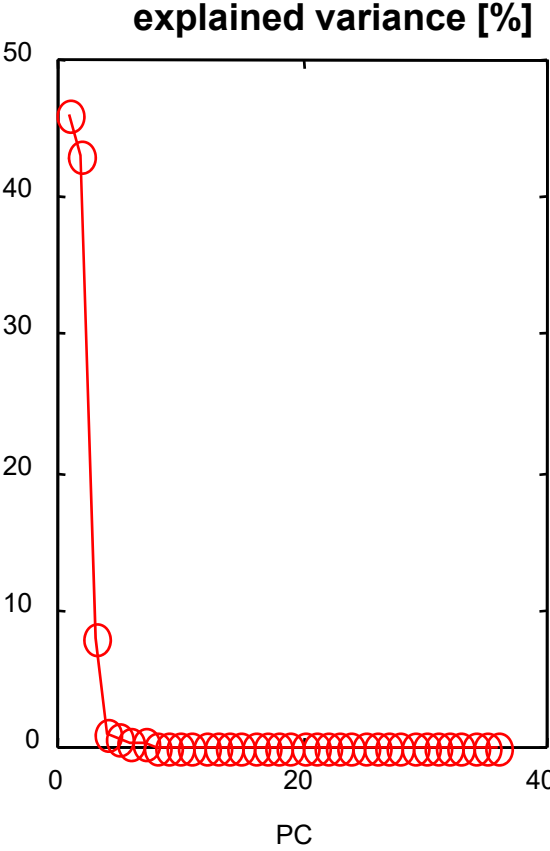
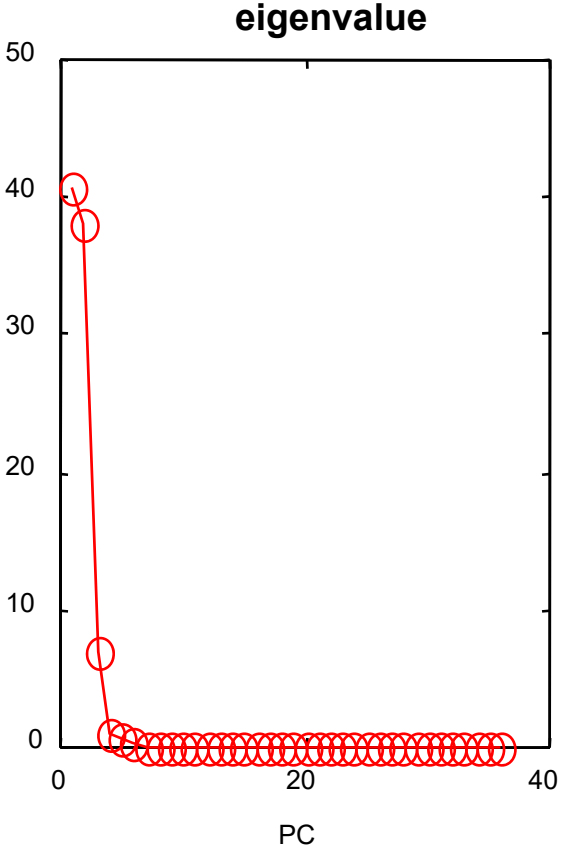
PCA computation

- The spectra average is reduced to zero therefore if the normal distribution assumption is satisfied, the whole information is in the covariance matrix.
- Eigenvectors and eigenvalues calculation

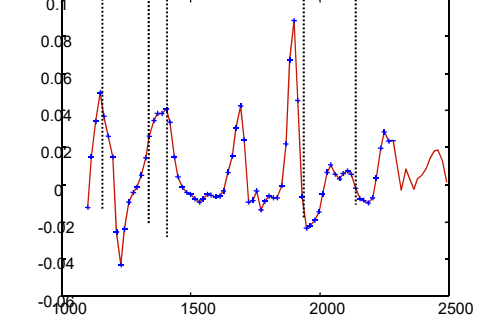
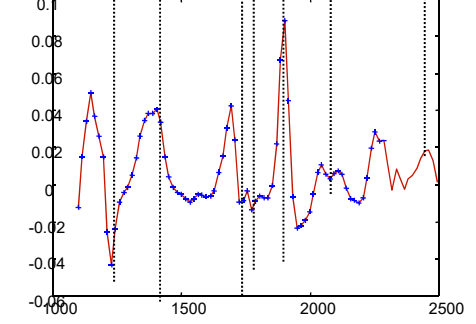
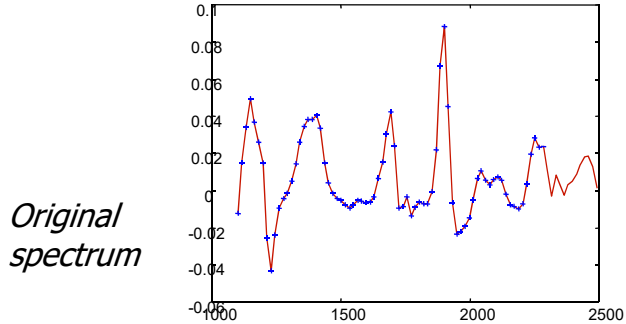
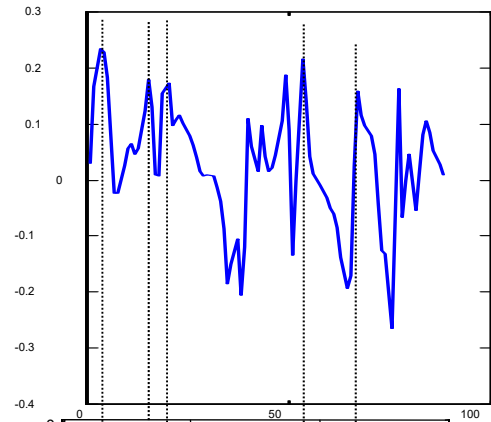
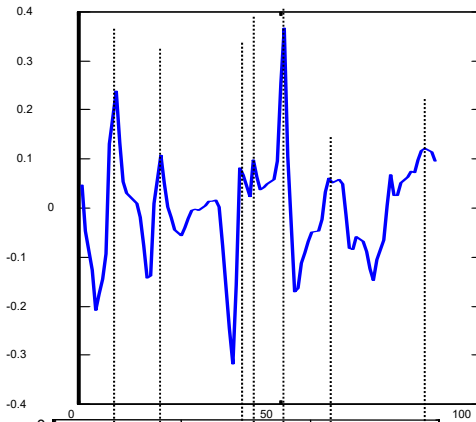
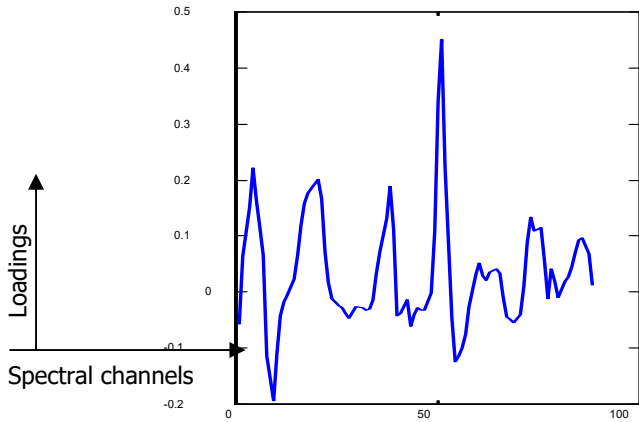
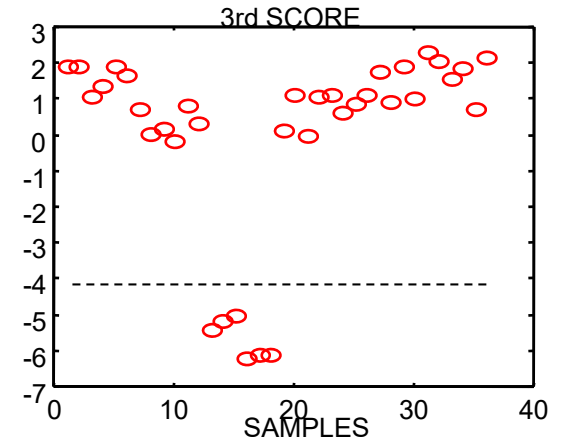
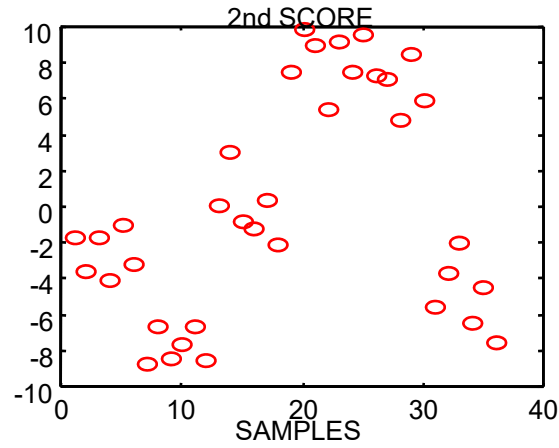
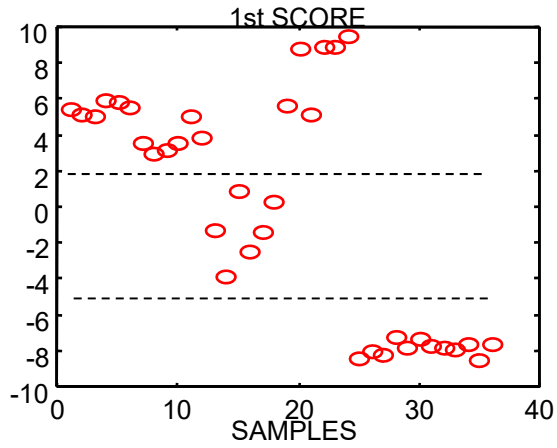


- The first 3 eigenvalues have values significantly different to zero.
- The 88 spectra, vectors in a dimensional space of 88, are largely limited to a subspace dimension of 3.

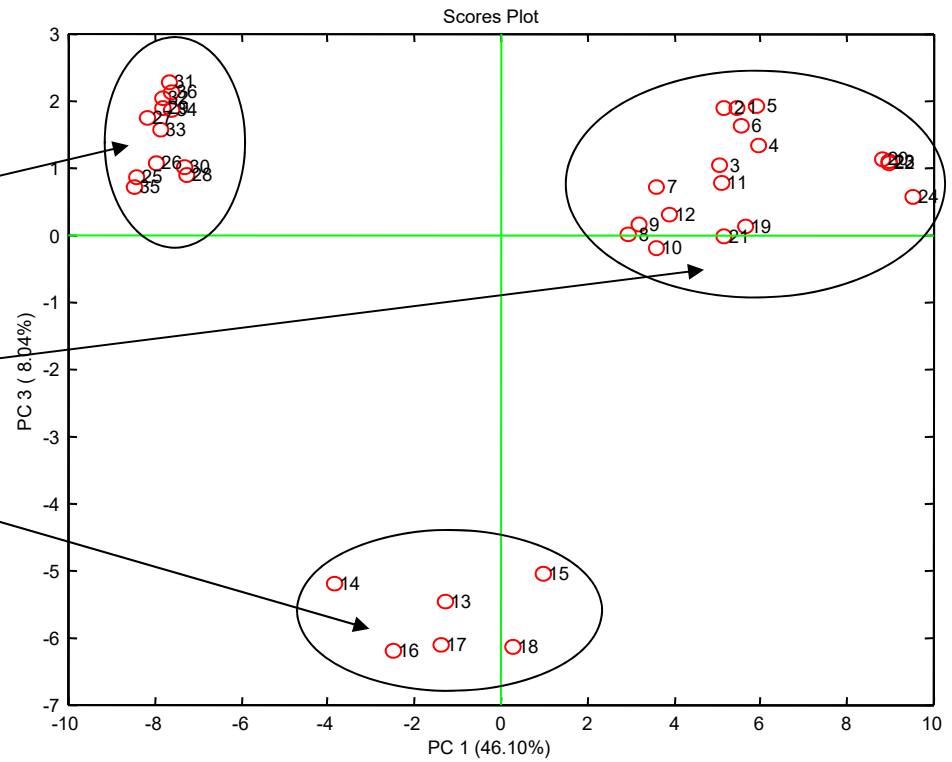
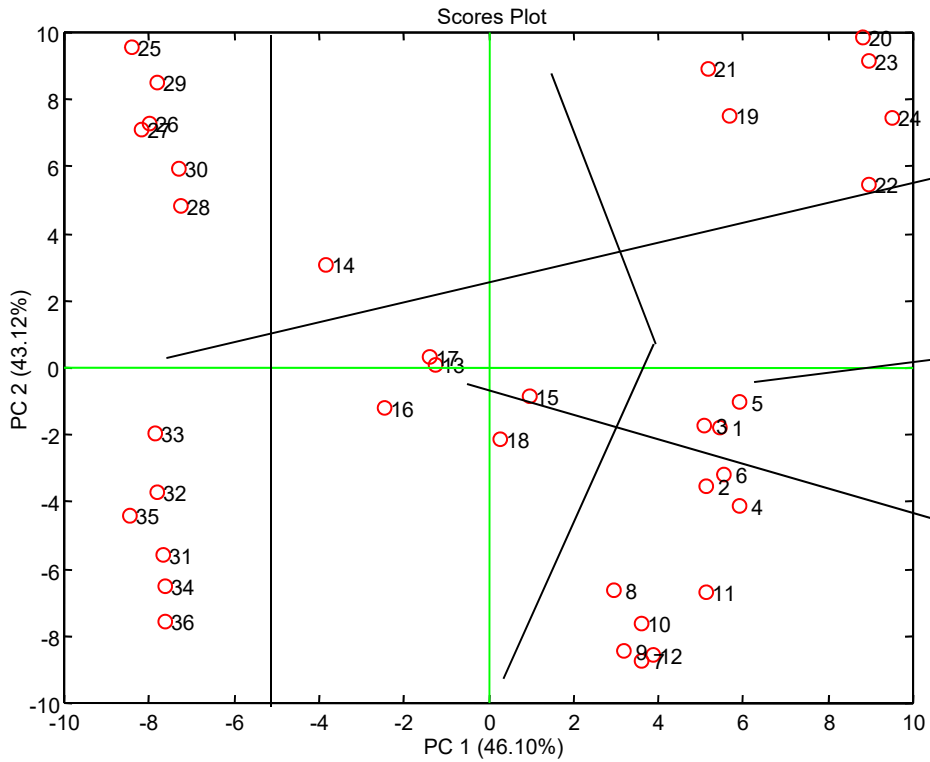
Eigenvalue and variance



Scores e loadings

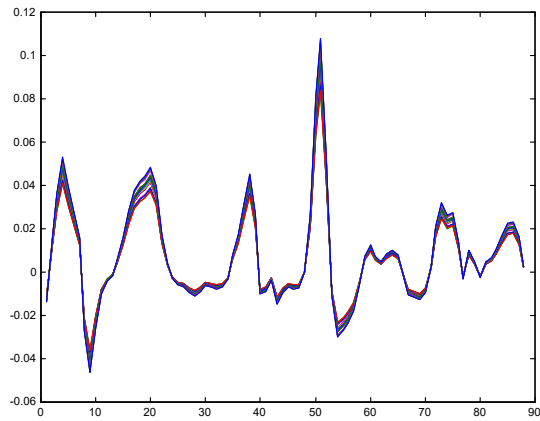


Scores plot

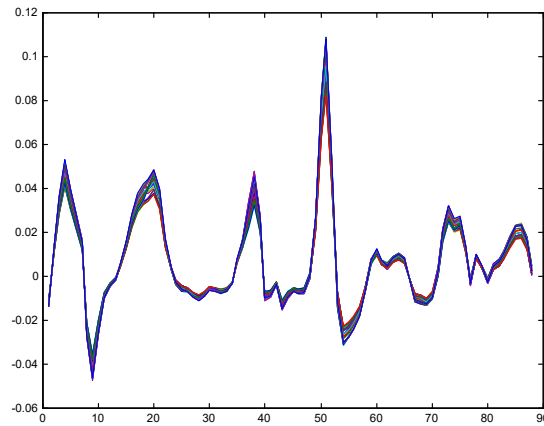


Decomposition and residues

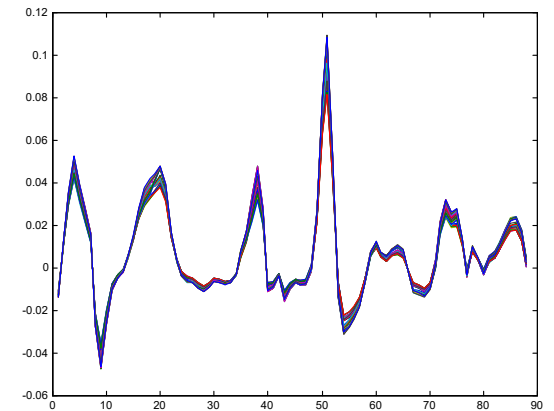
First PC



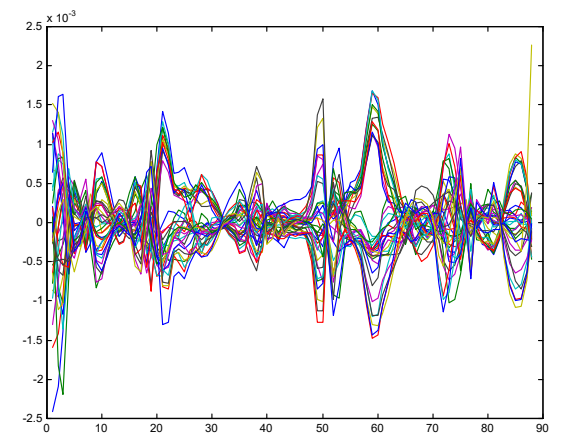
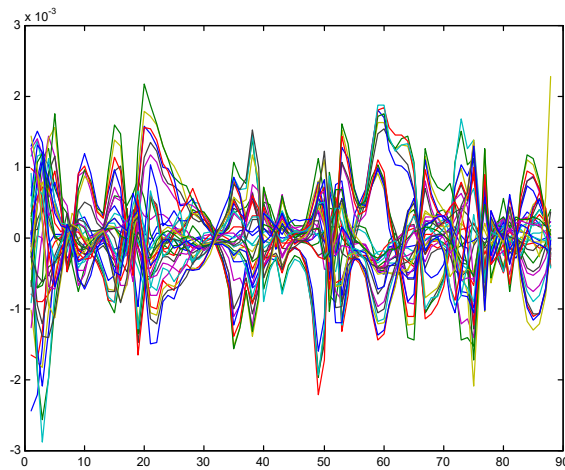
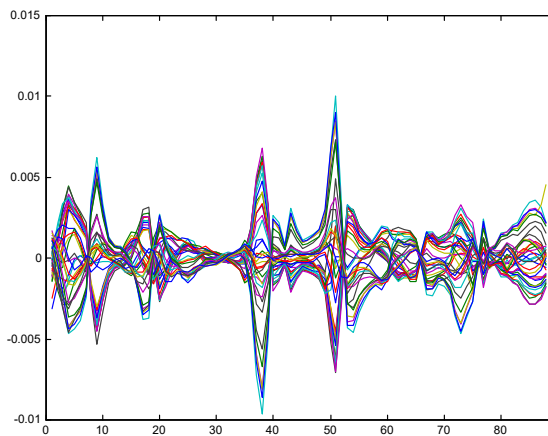
Second PC



Third PC



Residues



Principal Components Regression (PCR)

- We divide the dataset into two:
- 26 for the calculation of PCcal model, Y_{cal}
- 10 for the error evaluation PCval, Y_{val}
- The model calculates the regression matrix B_{pcr}

$$Y_{cal} = X_{cal} \cdot B^T \Rightarrow B^T = P \cdot \Lambda^{-1} \cdot T^T \cdot Y_{cal}$$

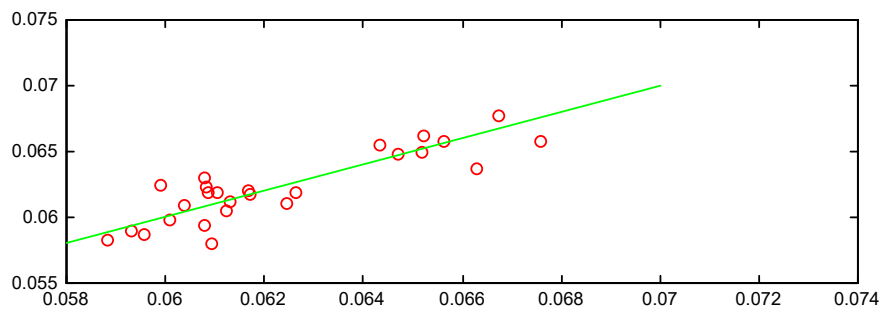
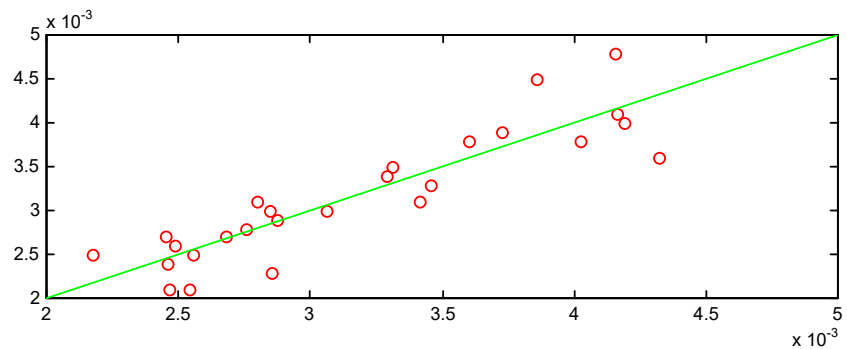
- We calculate an estimation of the validation set (and for comparison also of the calibration)
- RMSEC and RMSECV

$$stimaY_{cal} = X_{cal} \cdot B^T$$

$$stimaY_{val} = X_{val} \cdot B^T$$

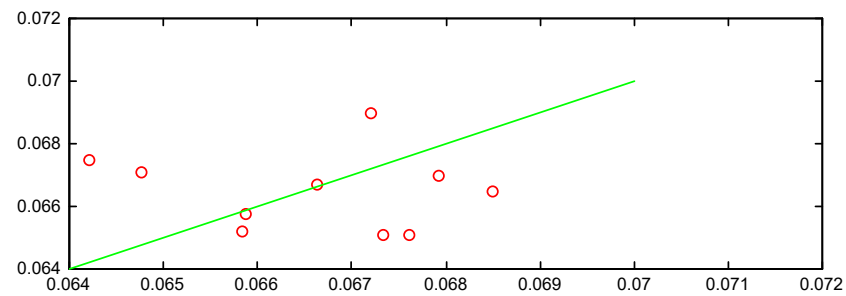
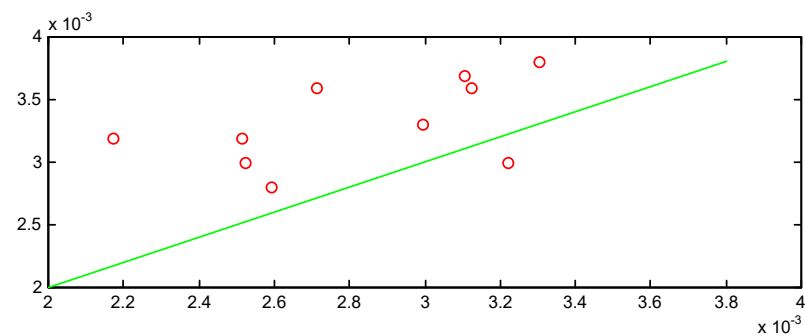
results

calibration



RMSEC_{acidity} = $3.1 \cdot 10^{-4}$
RMSEC_{humidity} = 0.0013

test



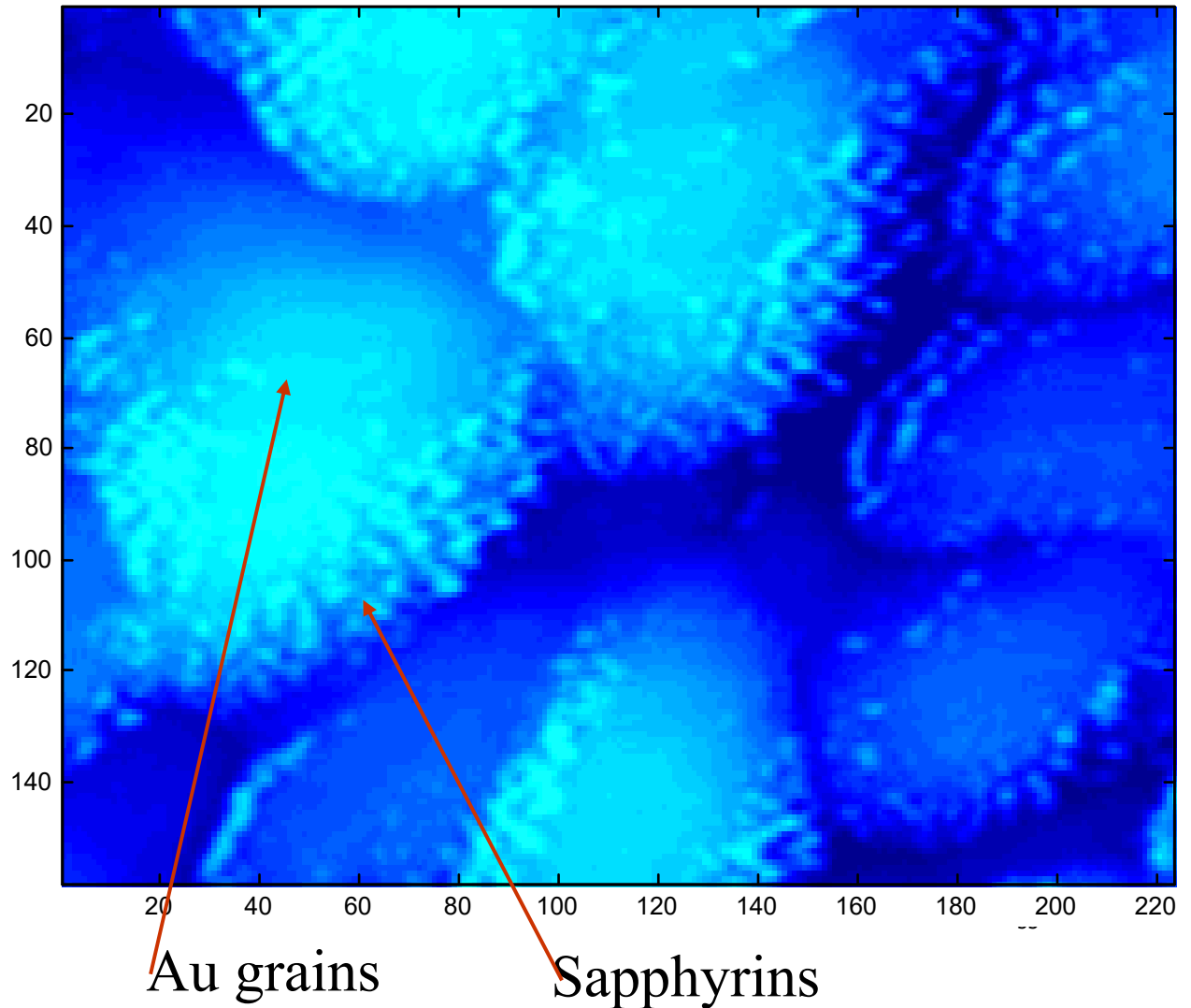
RMSECV_{acidity} = $5.9 \cdot 10^{-4}$
RMSECV_{humidity} = 0.0019

Application to the analysis of the images

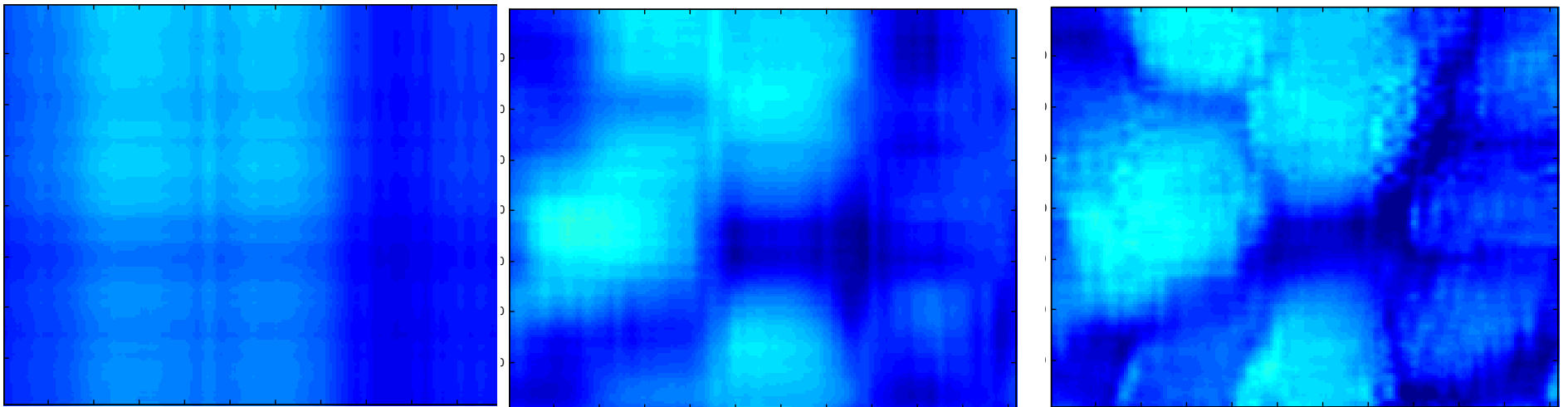
- A scanned image can be seen as an $N \times M$ matrix in the case of gray scale (black to white scale image) or $N \times M \times 3$ (in the case of color image)
- A picture can be considered as a matrix and we can apply the PCA
- The PCA decomposition may bring out some peculiar structures of the image allowing to study the characteristics of the image.

PCA: Application to Image Analysis (example 1: I)

- STM image of Sapphyrin molecules growth as a Langmuir-Blodgett film onto a gold substrate.



PCA: Application to Image Analysis (example 1: II)



$$X = S_1^T \cdot L_1$$

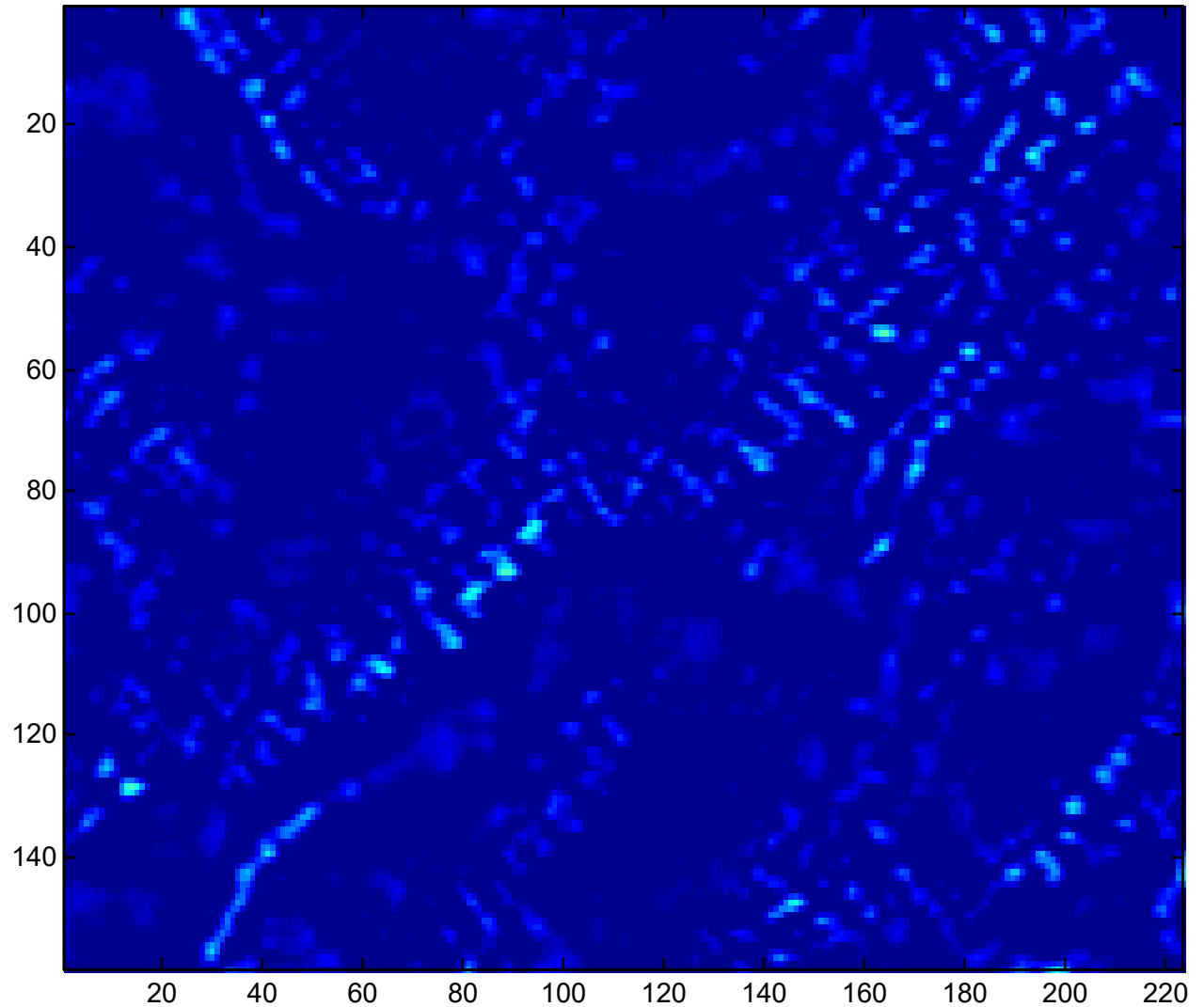
$$X = S_{1:10}^T \cdot L_{1:10}$$

$$X = S_{1:15}^T \cdot L_{1:15}$$

PCA: Application to Image Analysis (example 1: III)

- The residuals of the expansion at the tenth PC put in evidence the sapphyrine film only.

$$X - S_{1:10}^T \cdot L_{1:10}$$

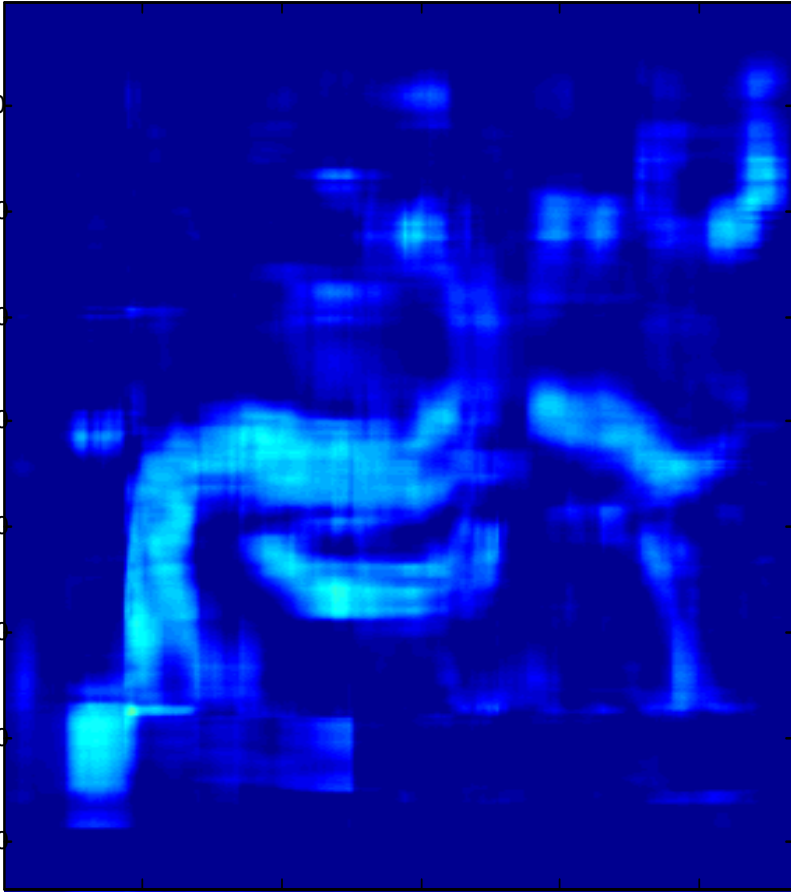


PCA: Application to Image Analysis (example 2: I)

- Caravaggio Deposition



PCA: Application to Image Analysis (example 2: II)



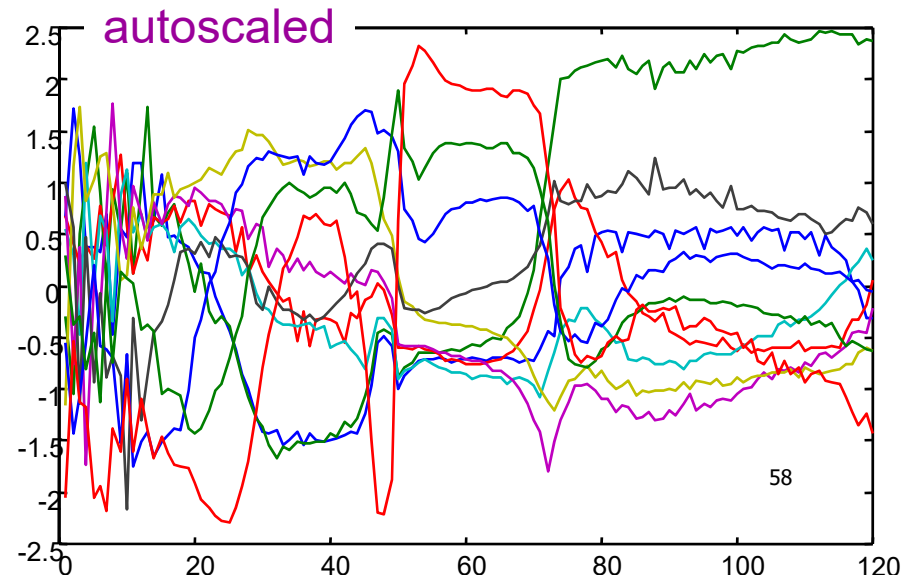
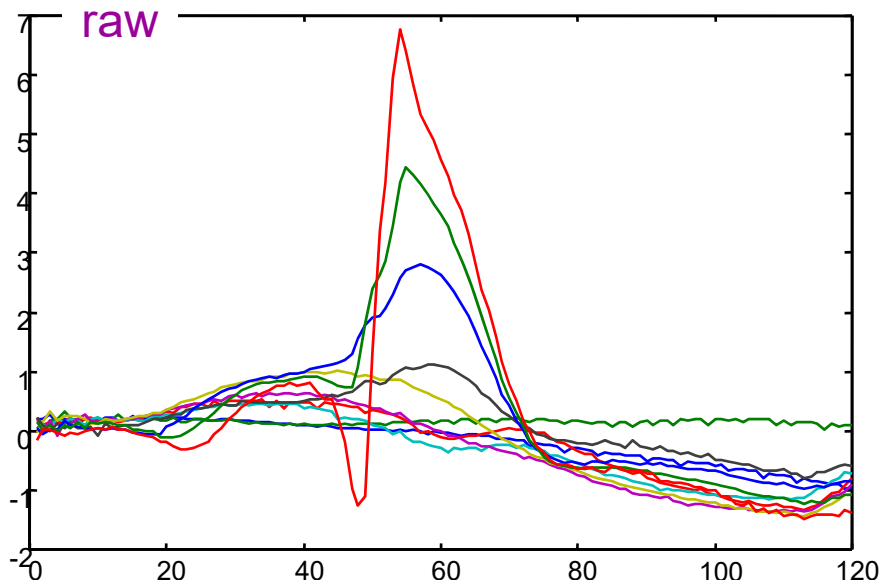
$$X = S_{1:10}^T \cdot L_{1:10}$$



$$X - S_{1:10}^T \cdot L_{1:10}$$

The normalization problem

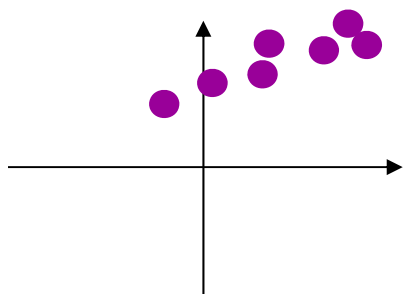
- The normalization is an operation that reduces the matrix columns (features) to zero average (zero average and variance equal to one).
- The autoscaling gives the same weight to every feature, this procedure is good if we are sure that every feature has the same importance in the problem.
- The autoscaling becomes dangerous when one or more features are noisy or when the numerical relationships between features are important
- Typical case is the spectroscopy where autoscaling completely destroys the information



Normalization and Pattern Recognition

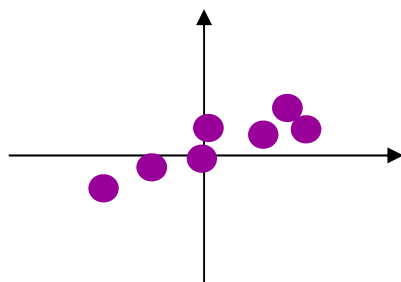
raw

X



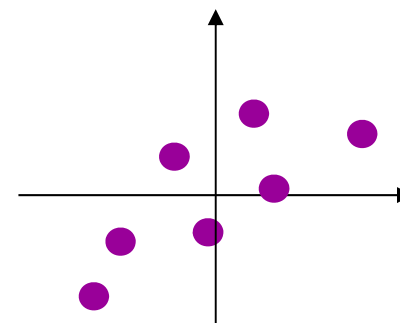
centered

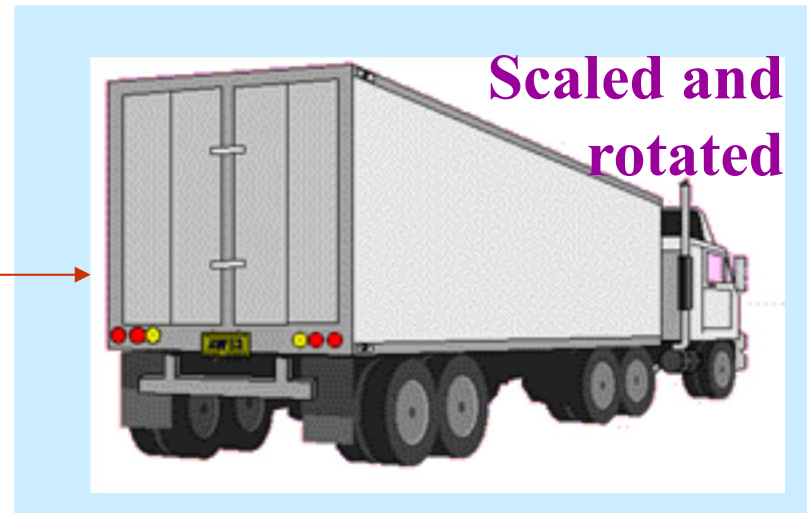
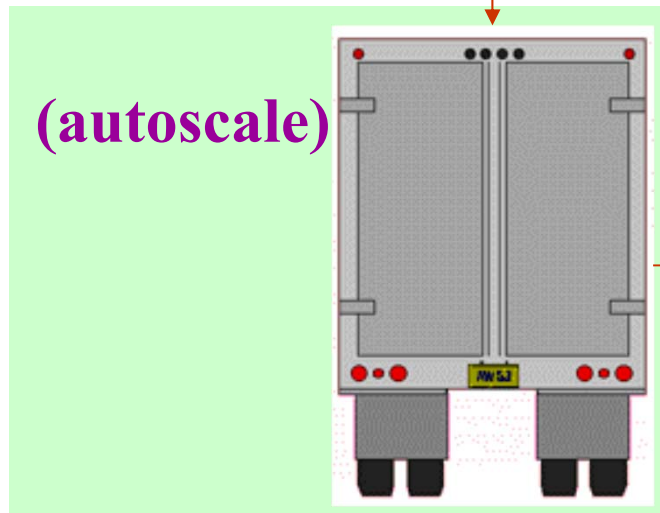
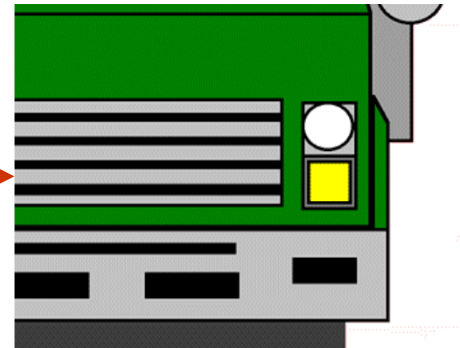
$G = X - \mu$



autoscaled

$Z = \frac{X - \mu}{\sigma}$





PCA and pattern recognition

- The principal component analysis is a method that allow:
- To define features of a new set (linear combination of the original) that are uncorrelated between them
- To decompose the variance of the data in the sum of the variance of the new axes (principal components)
- To reduce the representation of the pattern to a subspace identified by the main components of greatest variance
- To study the contribution of the original features to the core components by identifying the most significant higher contribution features.

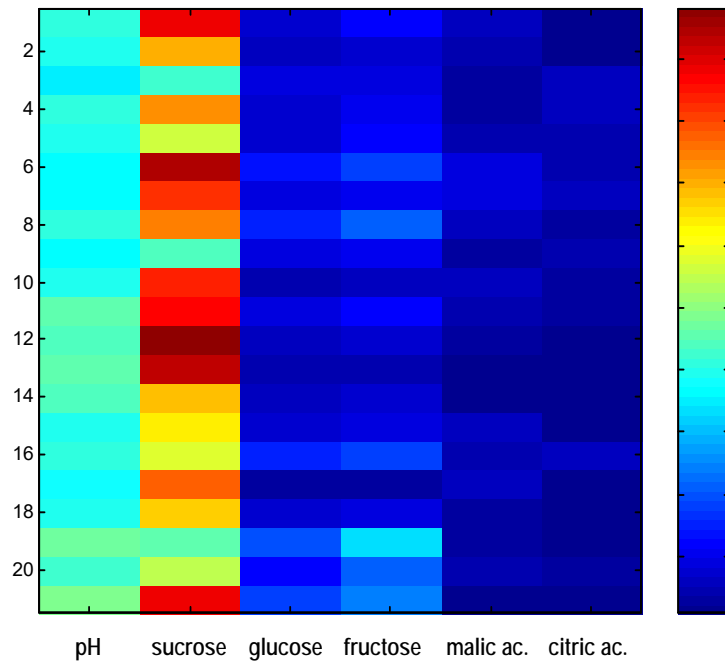
Example: Fruits parameters

- Suppose we have measured the following quantities in peaches: pH, sucrose, glucose, fructose, malic acid and citric acid, and we want to study the classification and the relationship using these parameters.

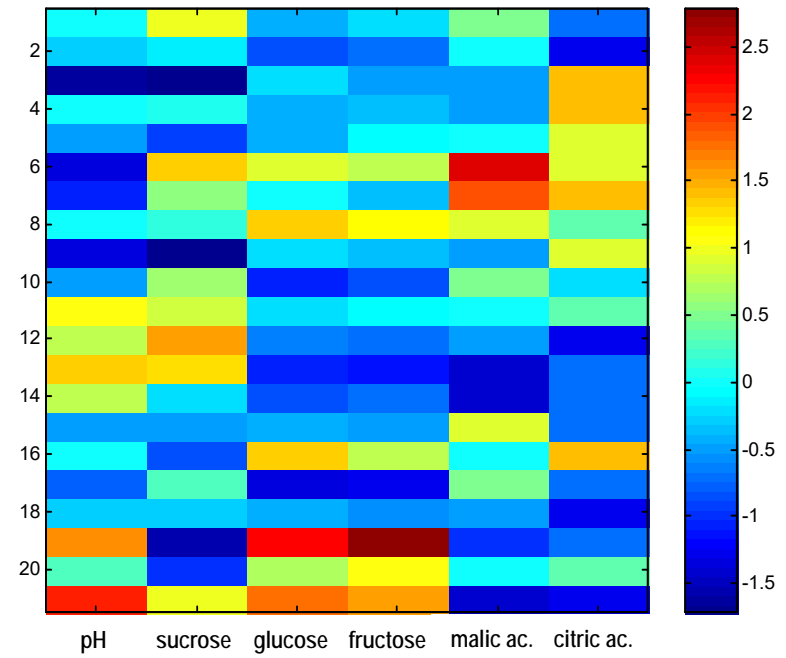
| | <i>pH</i> | <i>sucrose</i> | <i>glucose</i> | <i>fructose</i> | <i>malic acid</i> | <i>citric acid</i> |
|----------------------|-----------|----------------|----------------|-----------------|-------------------|--------------------|
| <i>baby gold</i> | 4.10 | 8.80 | 0.80 | 1.20 | 0.60 | 0.20 |
| <i>grezzano</i> | 4.0 | 7.0 | 0.60 | 0.80 | 0.50 | 0.10 |
| <i>iris rosso</i> | 3.50 | 4.30 | 0.90 | 1.0 | 0.40 | 0.60 |
| <i>maria aurelia</i> | 4.10 | 7.30 | 0.80 | 1.10 | 0.40 | 0.60 |
| <i>snow queen</i> | 3.90 | 5.70 | 0.80 | 1.30 | 0.50 | 0.50 |
| <i>spring star</i> | 3.60 | 9.40 | 1.40 | 1.90 | 1.0 | 0.50 |
| <i>super crimson</i> | 3.70 | 8.20 | 1.0 | 1.10 | 0.90 | 0.60 |
| <i>venus</i> | 4.10 | 7.40 | 1.60 | 2.20 | 0.70 | 0.40 |
| <i>argento roma</i> | 3.60 | 4.40 | 0.90 | 1.10 | 0.40 | 0.50 |
| <i>beauty lady</i> | 3.90 | 8.30 | 0.50 | 0.70 | 0.60 | 0.30 |
| <i>big top</i> | 4.50 | 8.60 | 0.90 | 1.30 | 0.50 | 0.40 |
| <i>doucer</i> | 4.40 | 9.80 | 0.70 | 0.80 | 0.40 | 0.10 |
| <i>felicia</i> | 4.60 | 9.30 | 0.50 | 0.50 | 0.20 | 0.20 |
| <i>kurakata</i> | 4.40 | 6.90 | 0.60 | 0.80 | 0.20 | 0.20 |
| <i>lucie</i> | 3.90 | 6.40 | 0.80 | 1.0 | 0.70 | 0.20 |
| <i>morsinai</i> | 4.10 | 5.80 | 1.60 | 1.90 | 0.50 | 0.60 |
| <i>oro</i> | 3.80 | 7.70 | 0.40 | 0.40 | 0.60 | 0.20 |
| <i>royal glory</i> | 4.0 | 6.70 | 0.80 | 0.90 | 0.40 | 0.10 |
| <i>sensation</i> | 4.70 | 4.60 | 2.0 | 3.40 | 0.30 | 0.20 |
| <i>sweet lady</i> | 4.20 | 5.50 | 1.30 | 2.10 | 0.50 | 0.40 |
| <i>youyeong</i> | 4.90 | 8.80 | 1.80 | 2.50 | 0.20 | 0.10 |

Color map

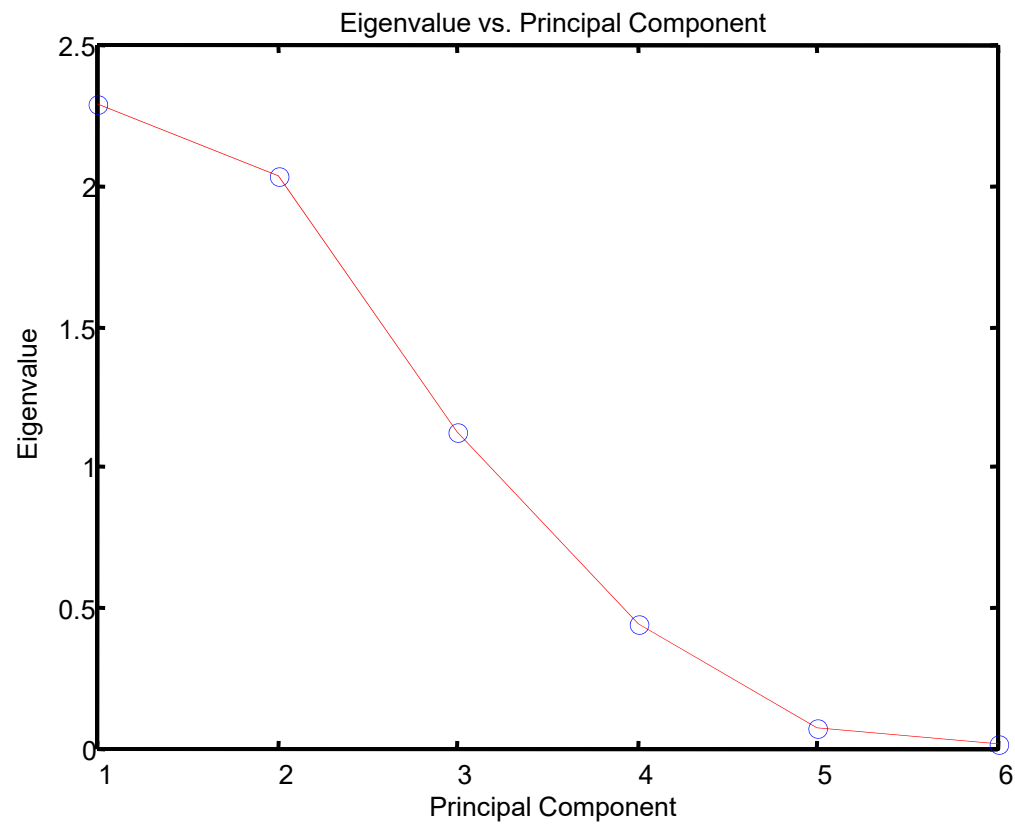
raw data



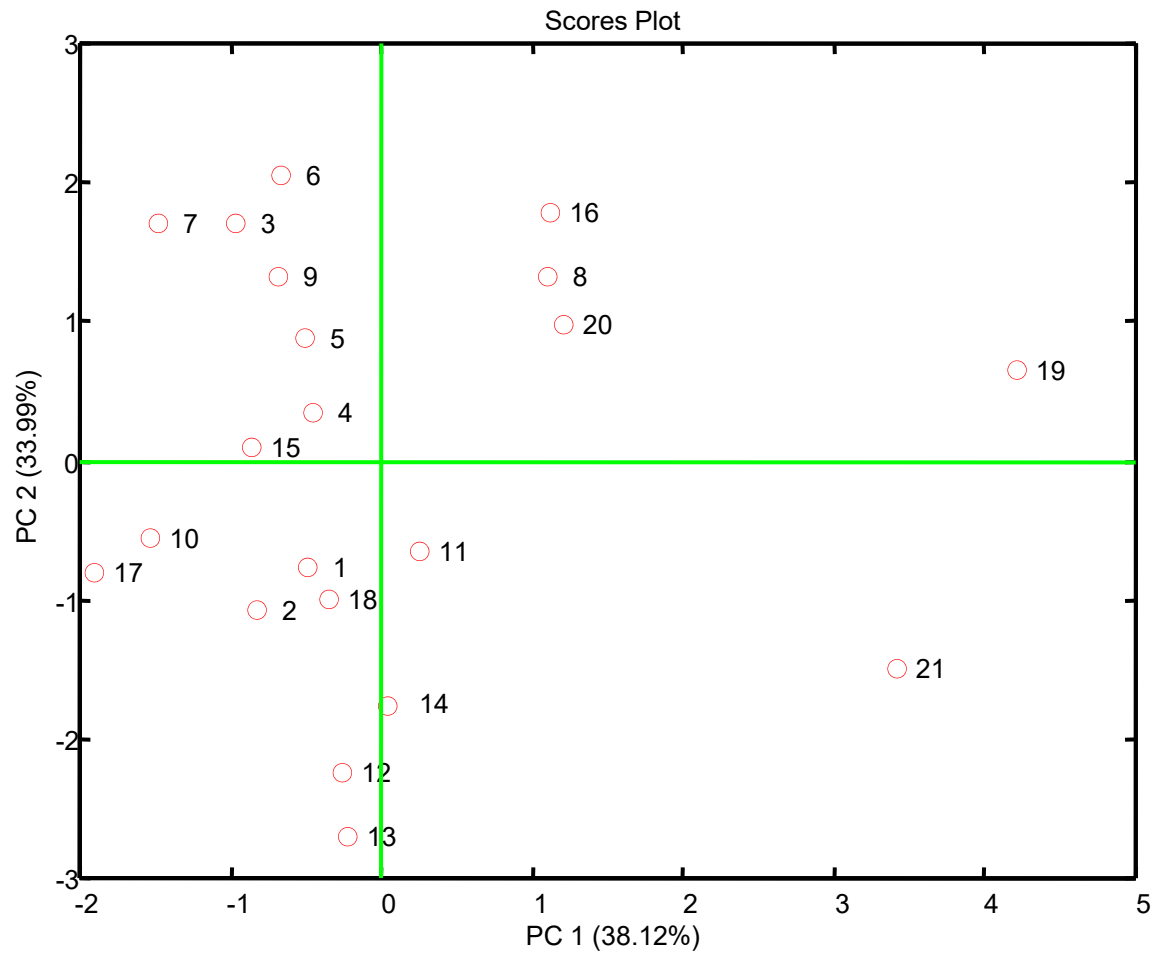
Autoscaled data



PCA: peaches data eigenvalues vs. PC



PCA: scores plot

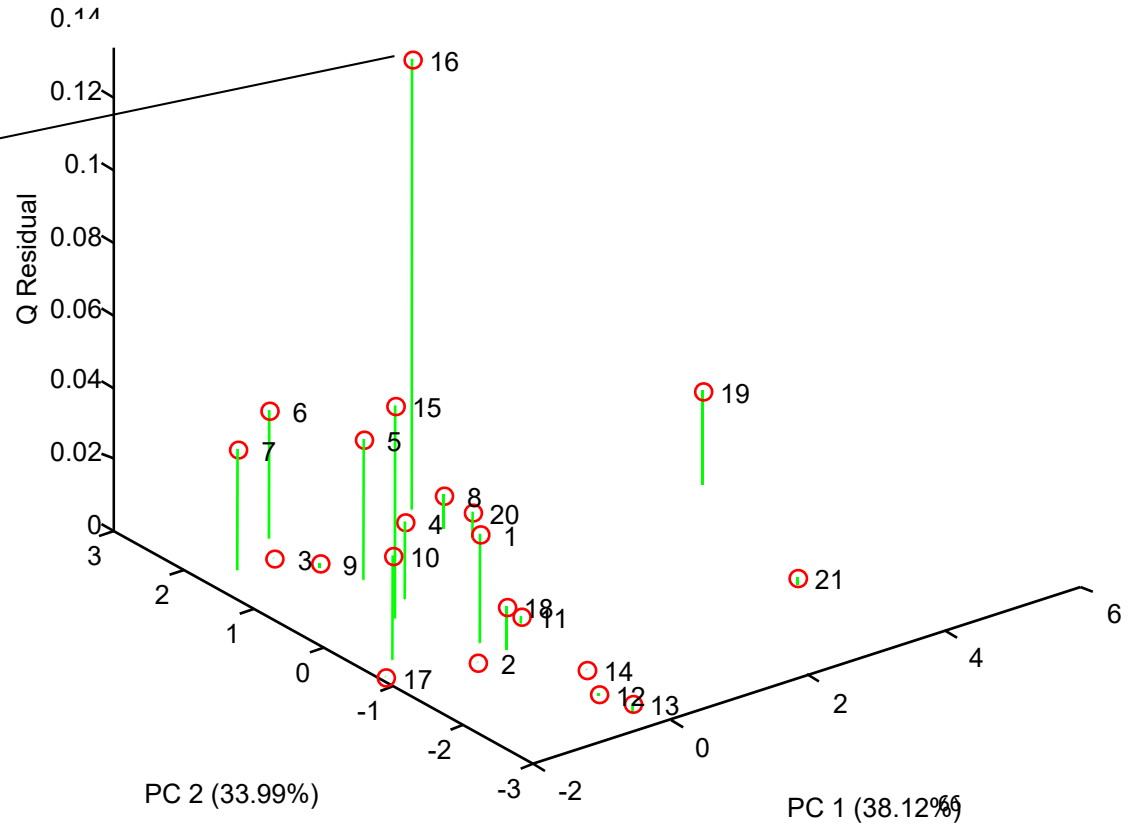
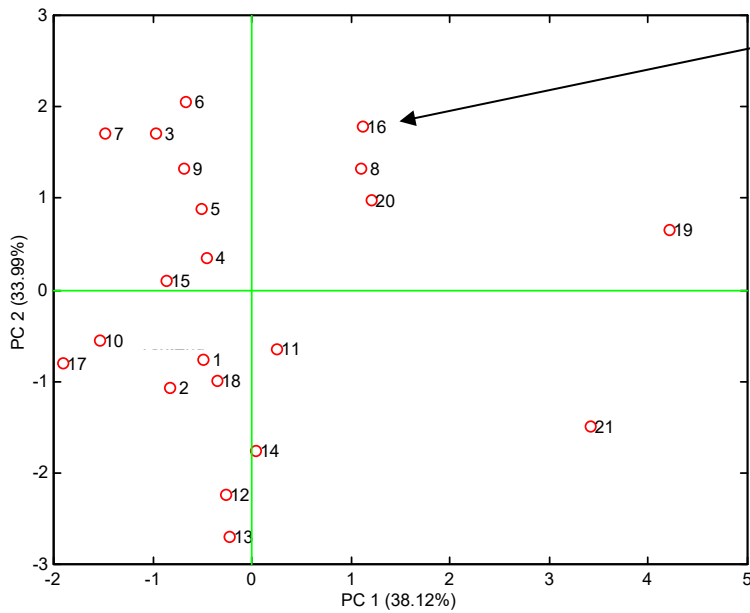


Residuals representation

$$x_i = a \cdot s_1 + b \cdot s_2 + \dots + n \cdot s_n$$

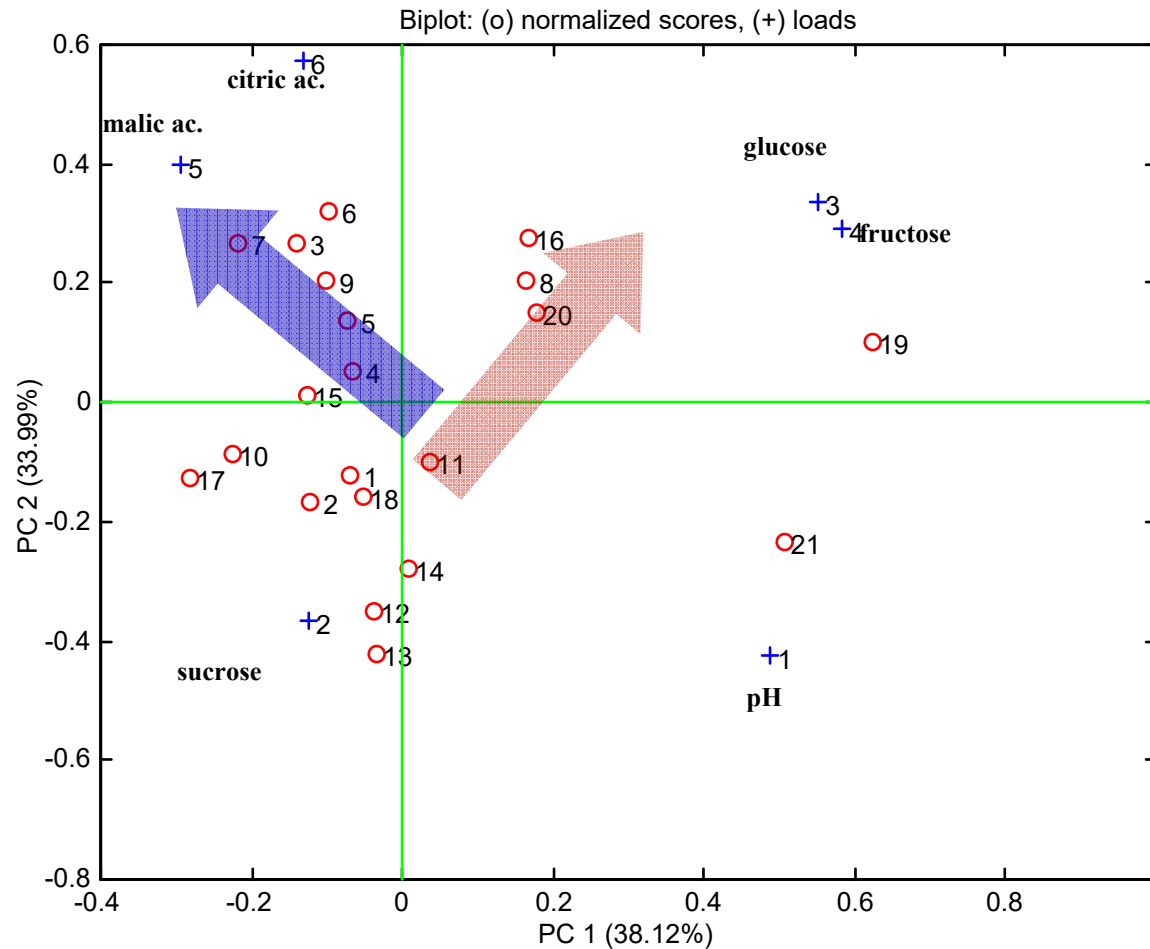
$$x_i^{pca} = a \cdot pc_1 + b \cdot pc_2 + residual$$

Scores Plot



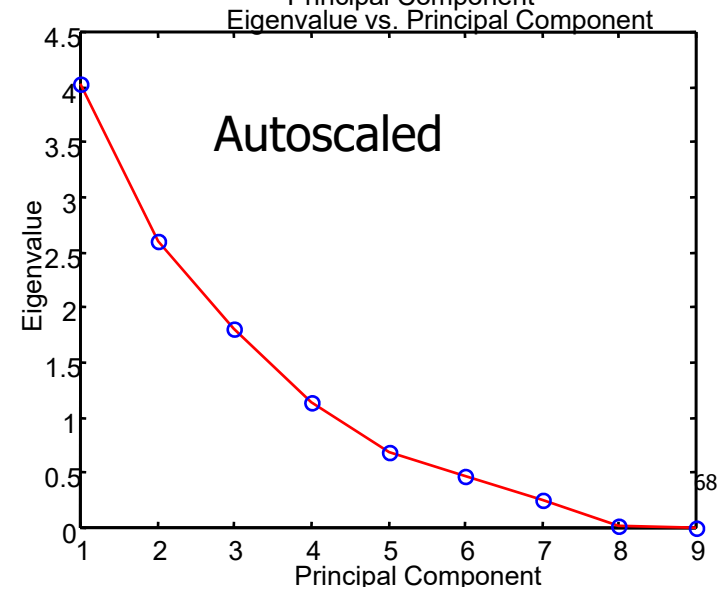
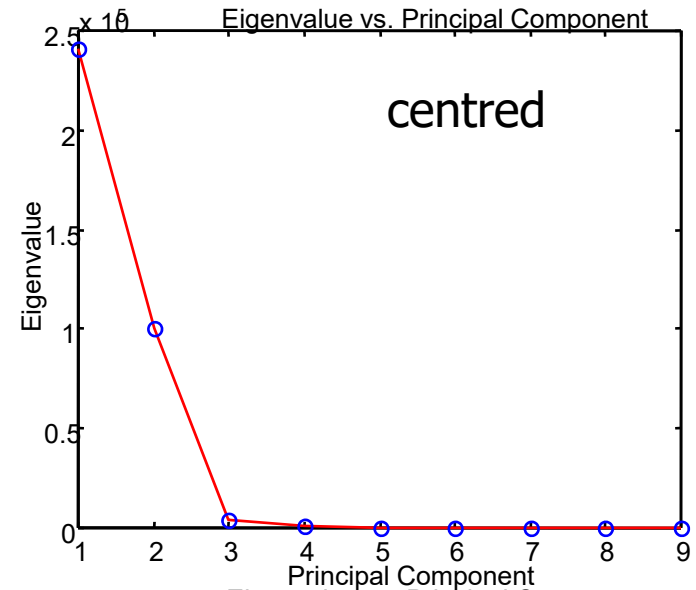
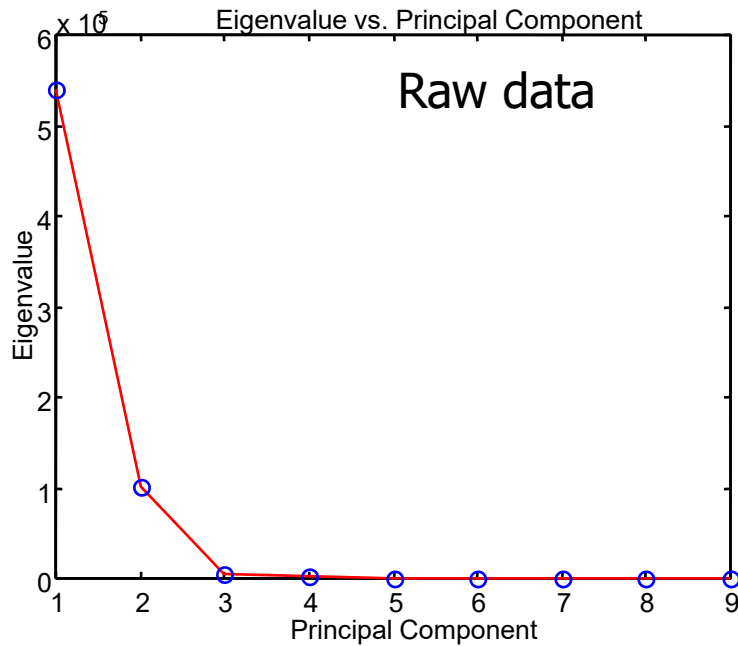
PCA example: peaches data

bi-plot: scores+loadings



- The sugars are orthogonal to acids
- We identify the direction of the acidity and the sweetness
- The sucrose is anticorrelated to glucose and fructose
- The pH is obviously anticorrelated to acids

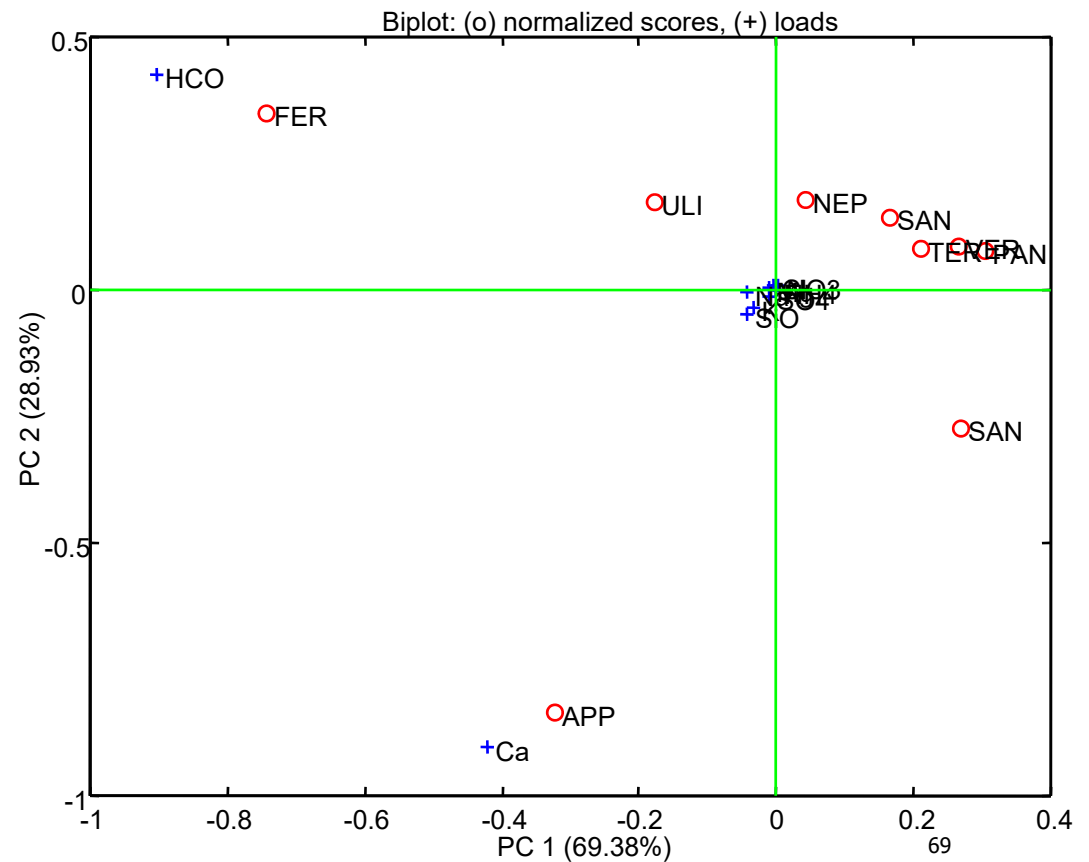
PCA example: mineral waters



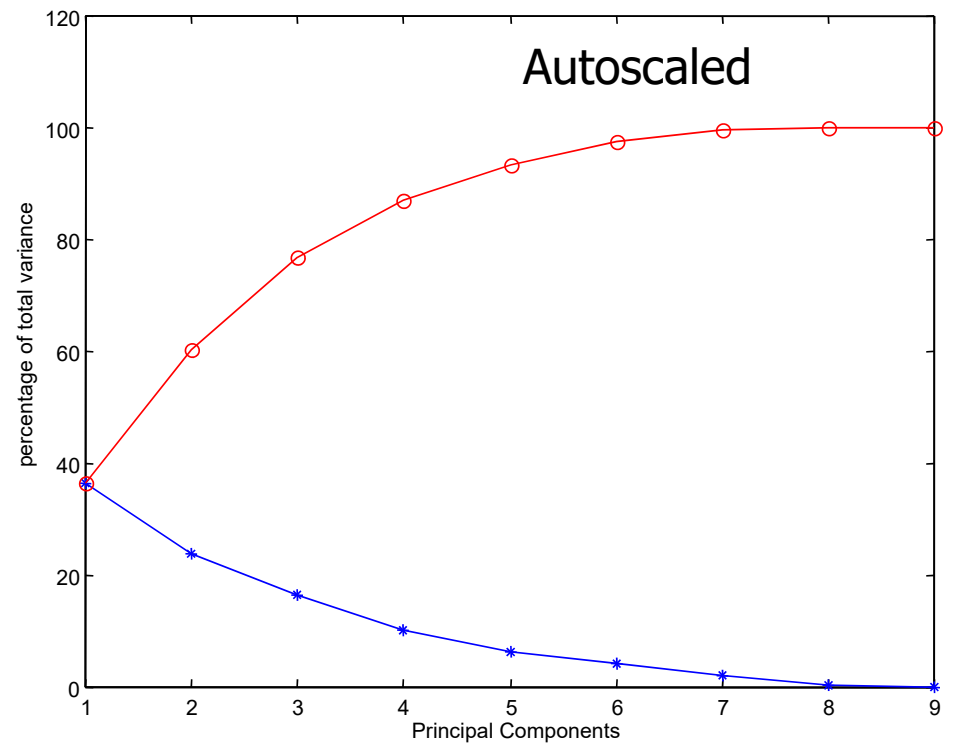
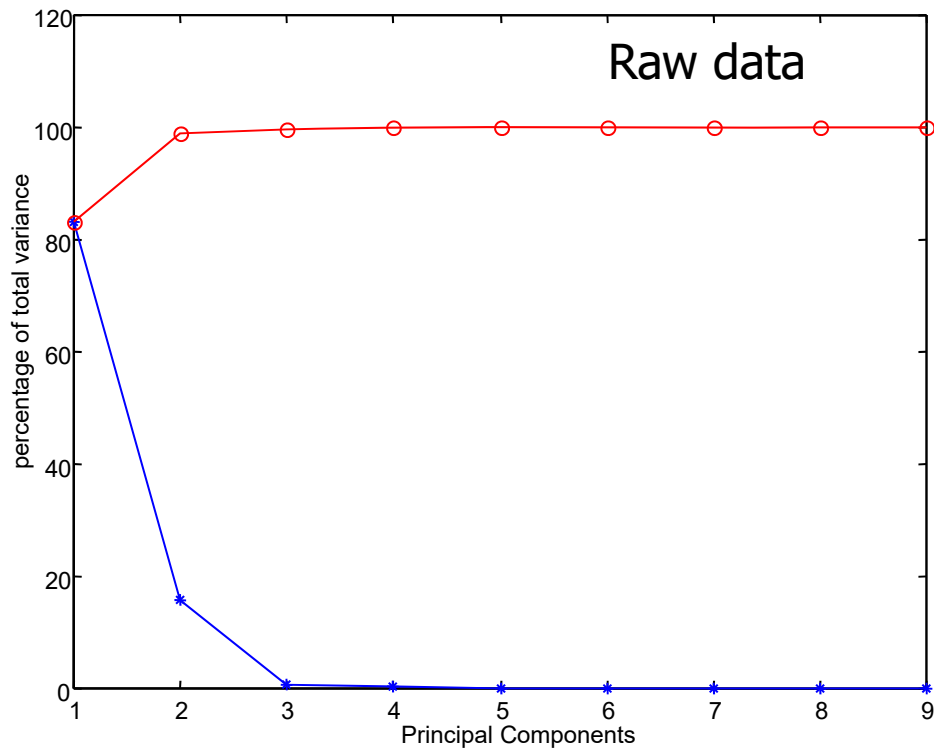
The autoscaling makes homogeneous the features by increasing the number of important dimensions

Mineral waters: PCA biplot raw data

- Only features numerically significant are important (HCO and Ca)
- The other features are around the origin and don't contribute to the classification
- HCO and Ca are orthogonal
- Orthogonal means uncorrelated
- Only RES and APP are different from others
- in this plot has 98% of the variance

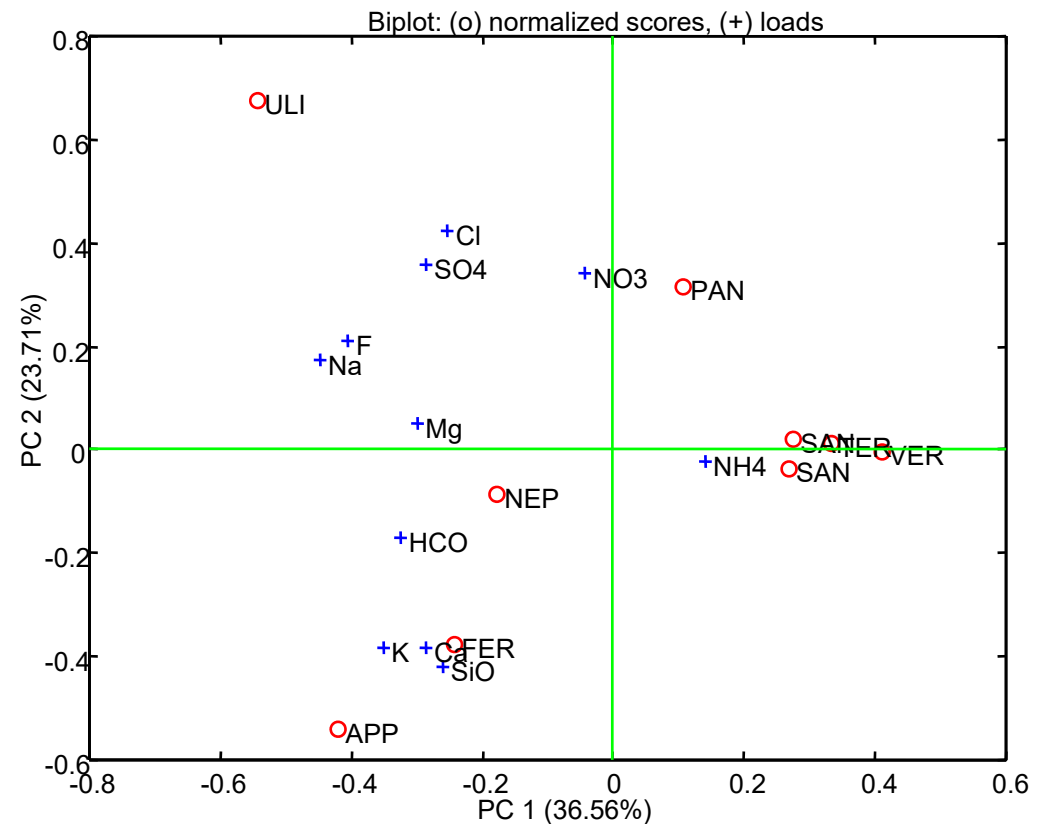


Scree plot variance

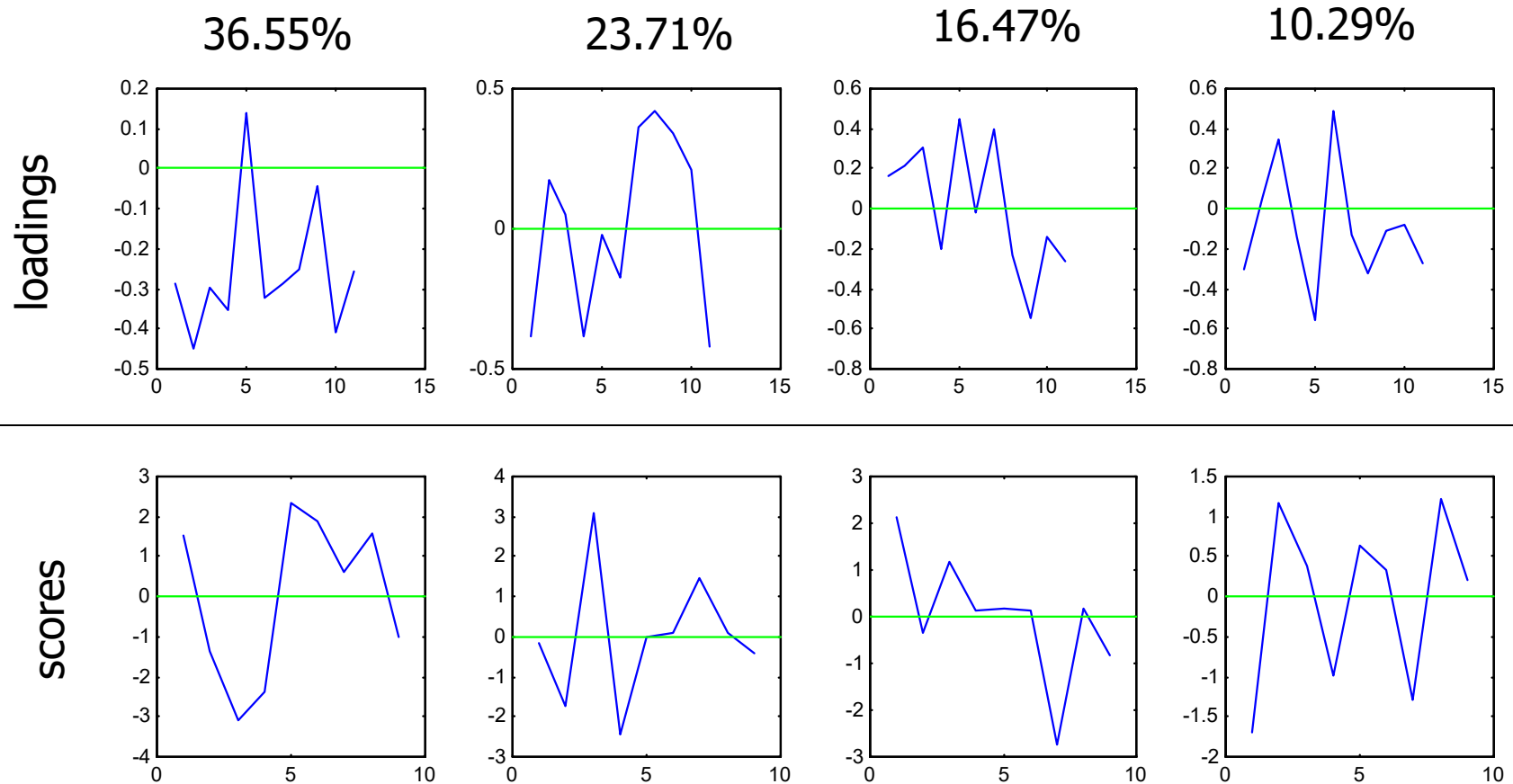


Mineral waters: Autoscaled PCA biplot

- The features contribute homogeneously
- The following are groups of waters:
 - SAN, TER, VER, SAB
 - minerals oligo
 - PAN
 - Oligo but with increase of NO₃
 - NEP, FER, APP
 - Intensification of Mg, HCO, Ca, K
 - ULI
 - Increasing in Cl, SO₄
 - For ULI, NEP, FER, APP
 - Common increasing of F, Na
 - 60% of the variance in this plot
 - And the other 40%?



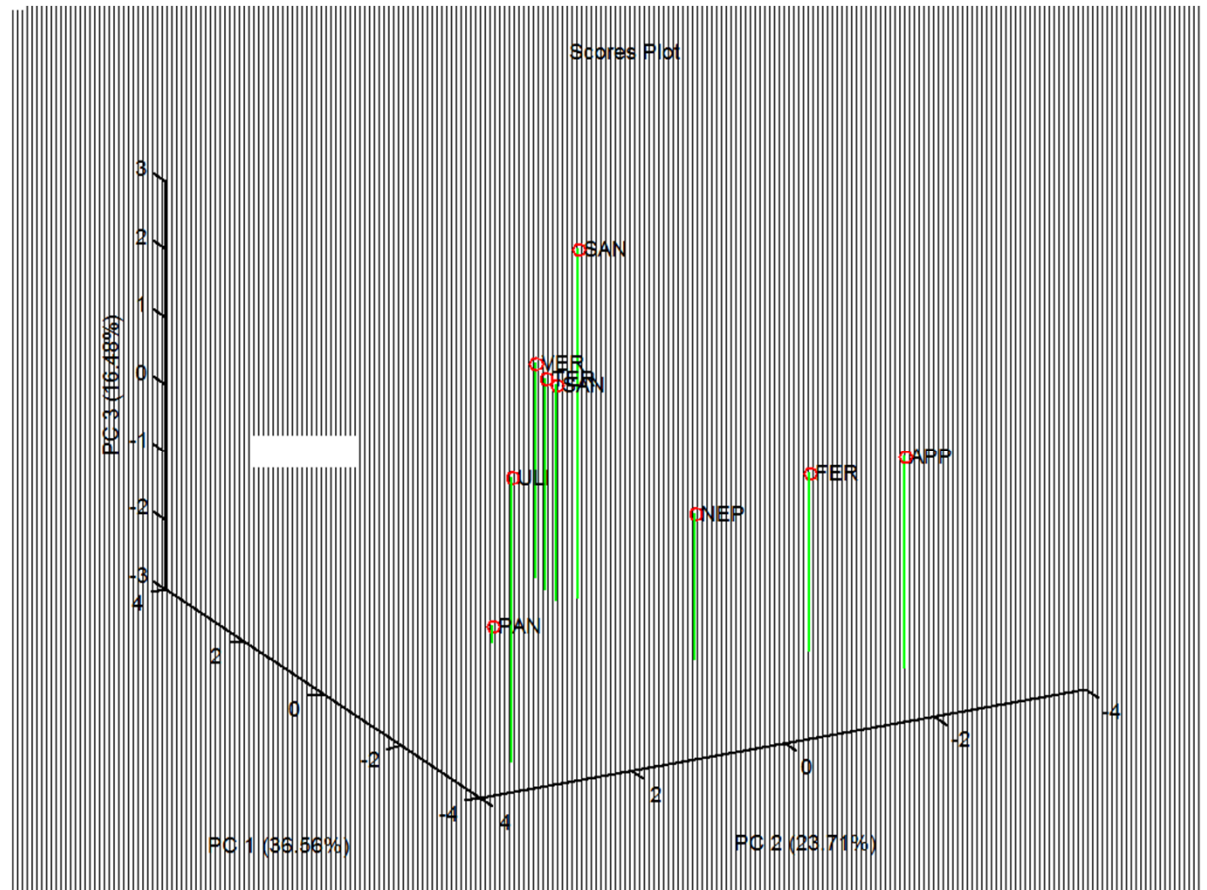
Mineral waters: loadings and scores analysis



PCA: mineral waters

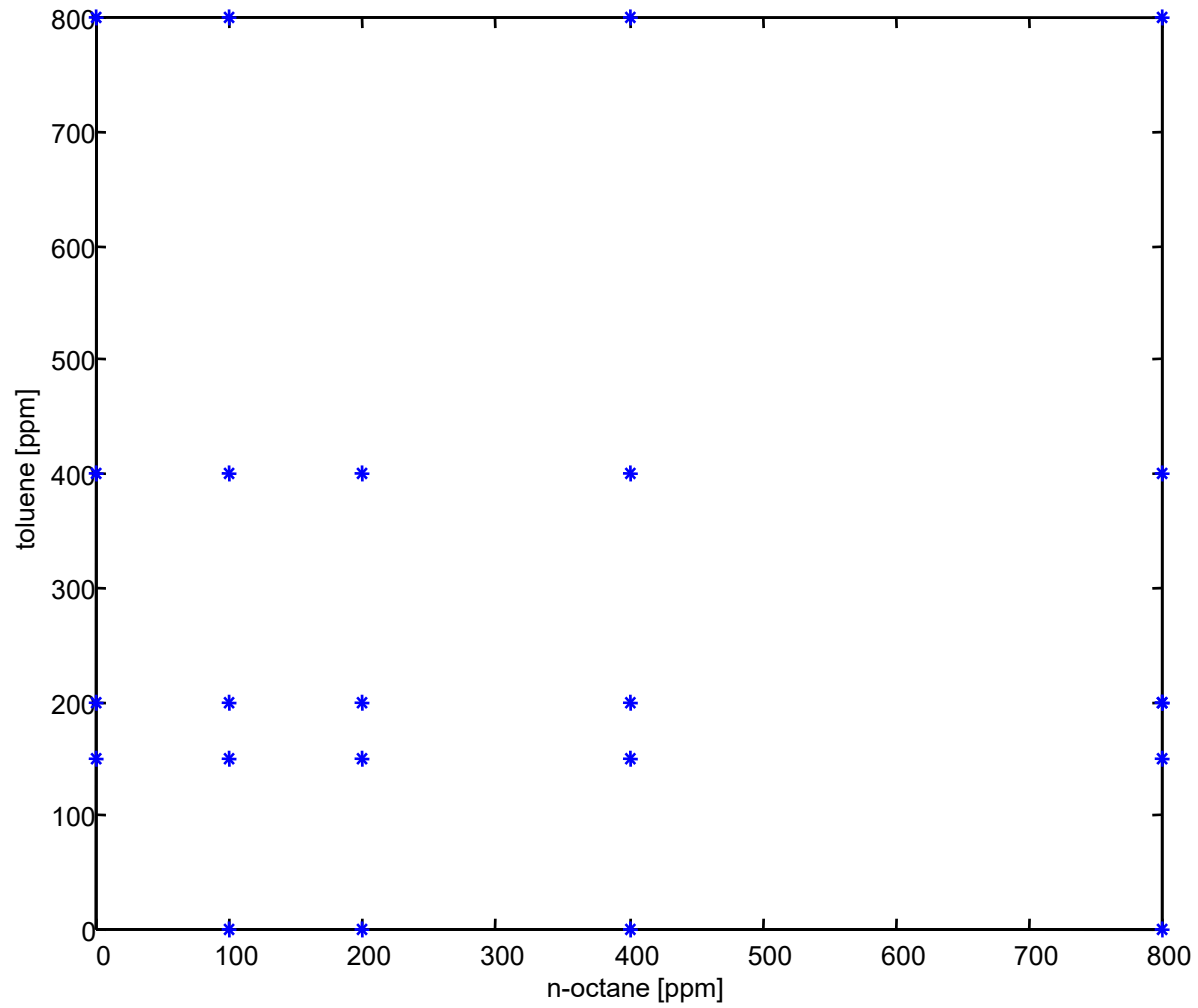
- The 2D representation is not sufficient because the distribution of eigenvalues. 2D representations capture only different aspects of the problem.

Score plot 3D
76% di variance
SAN is separated showing
specific characteristics

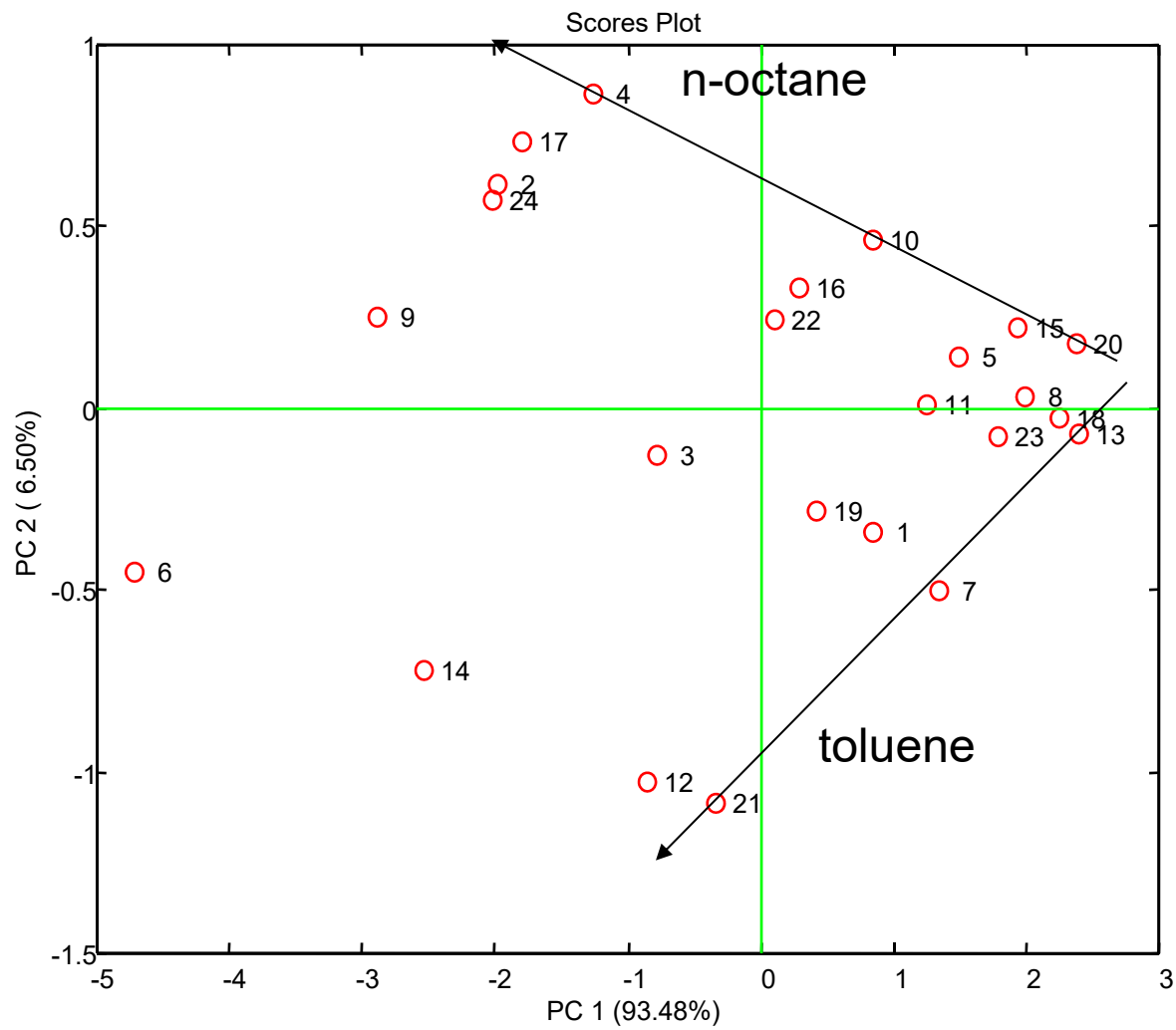


4 sensors for 2 gas

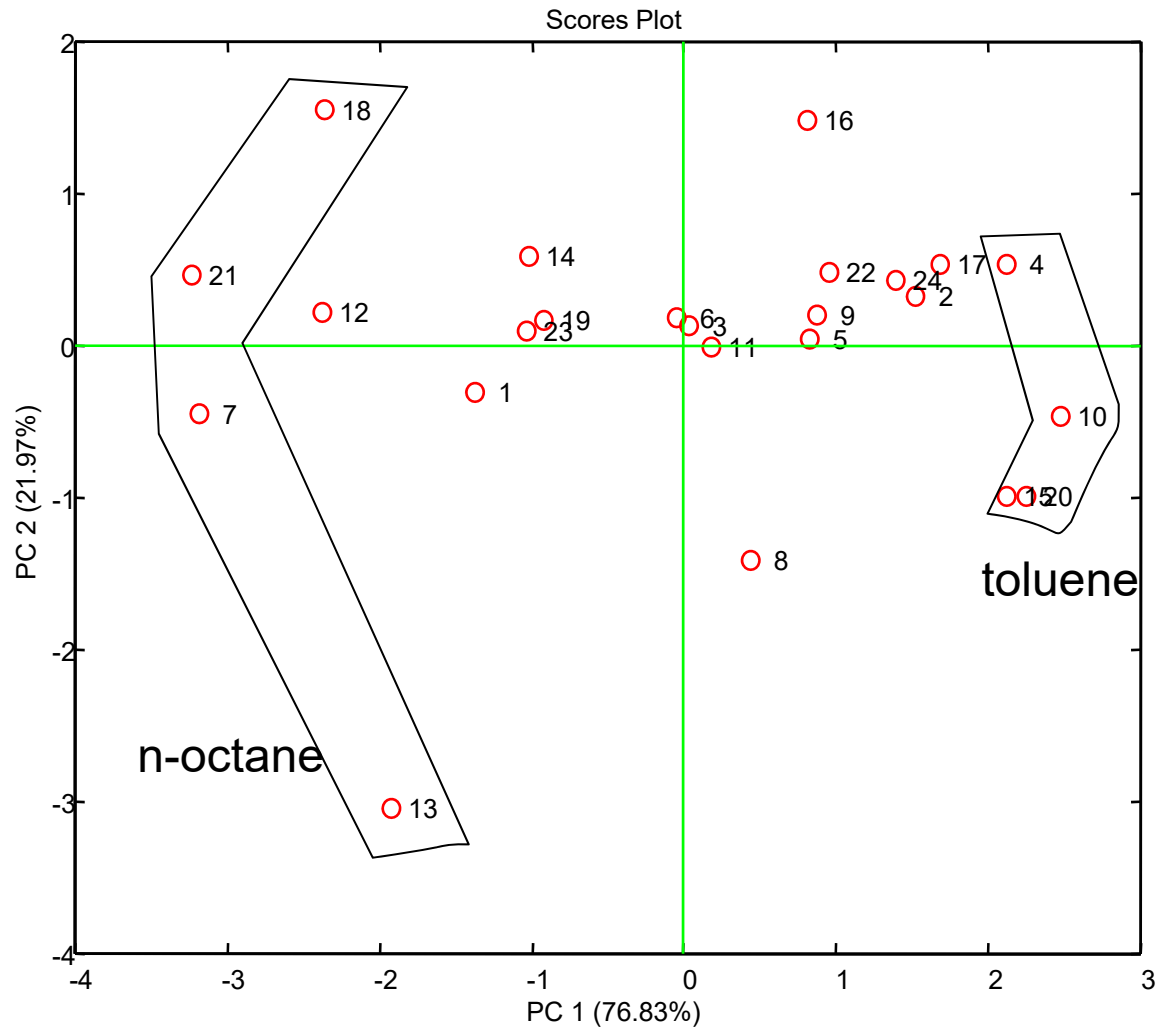
Measurements Plot



PCA scores

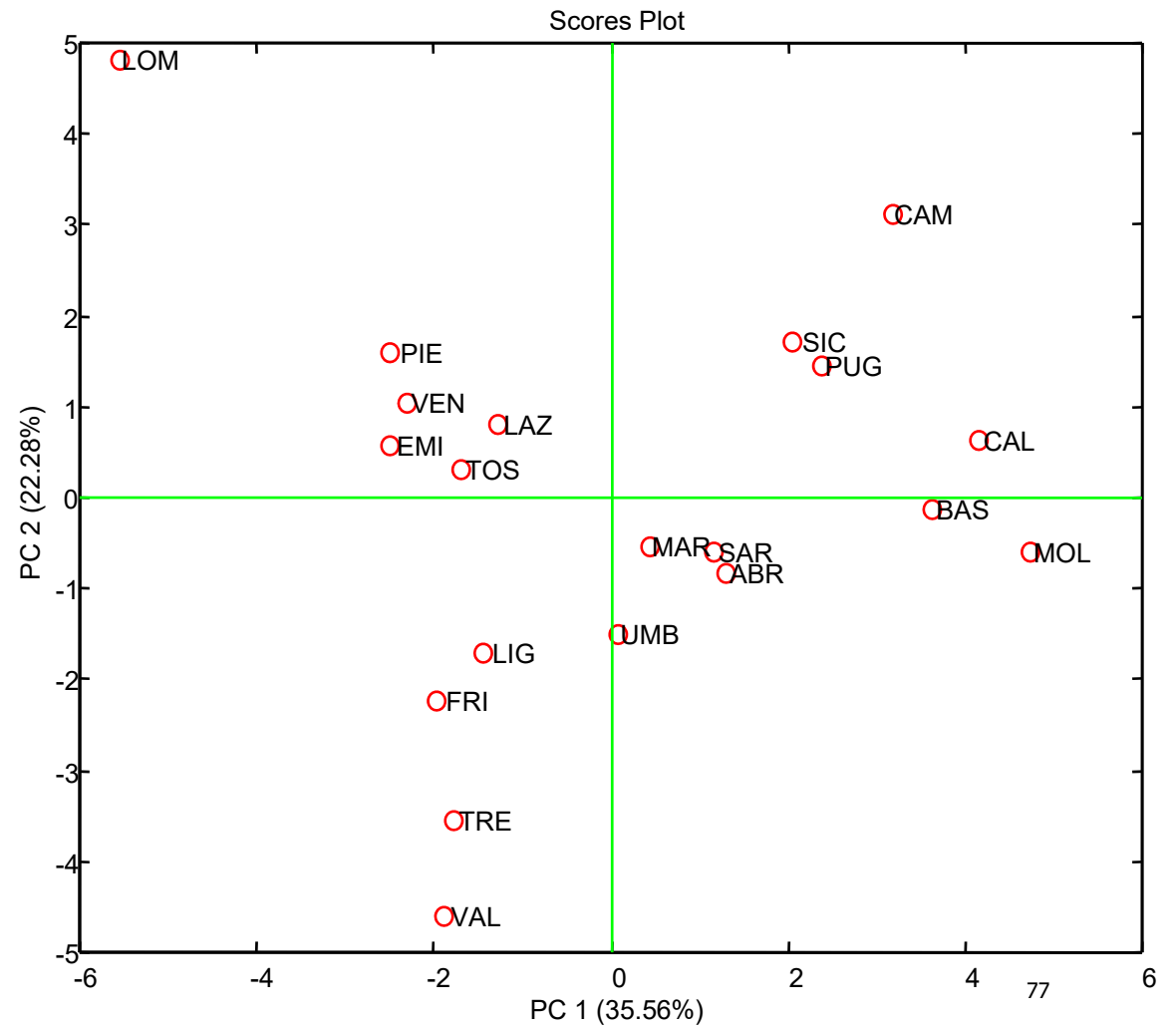


PCA scores normalization

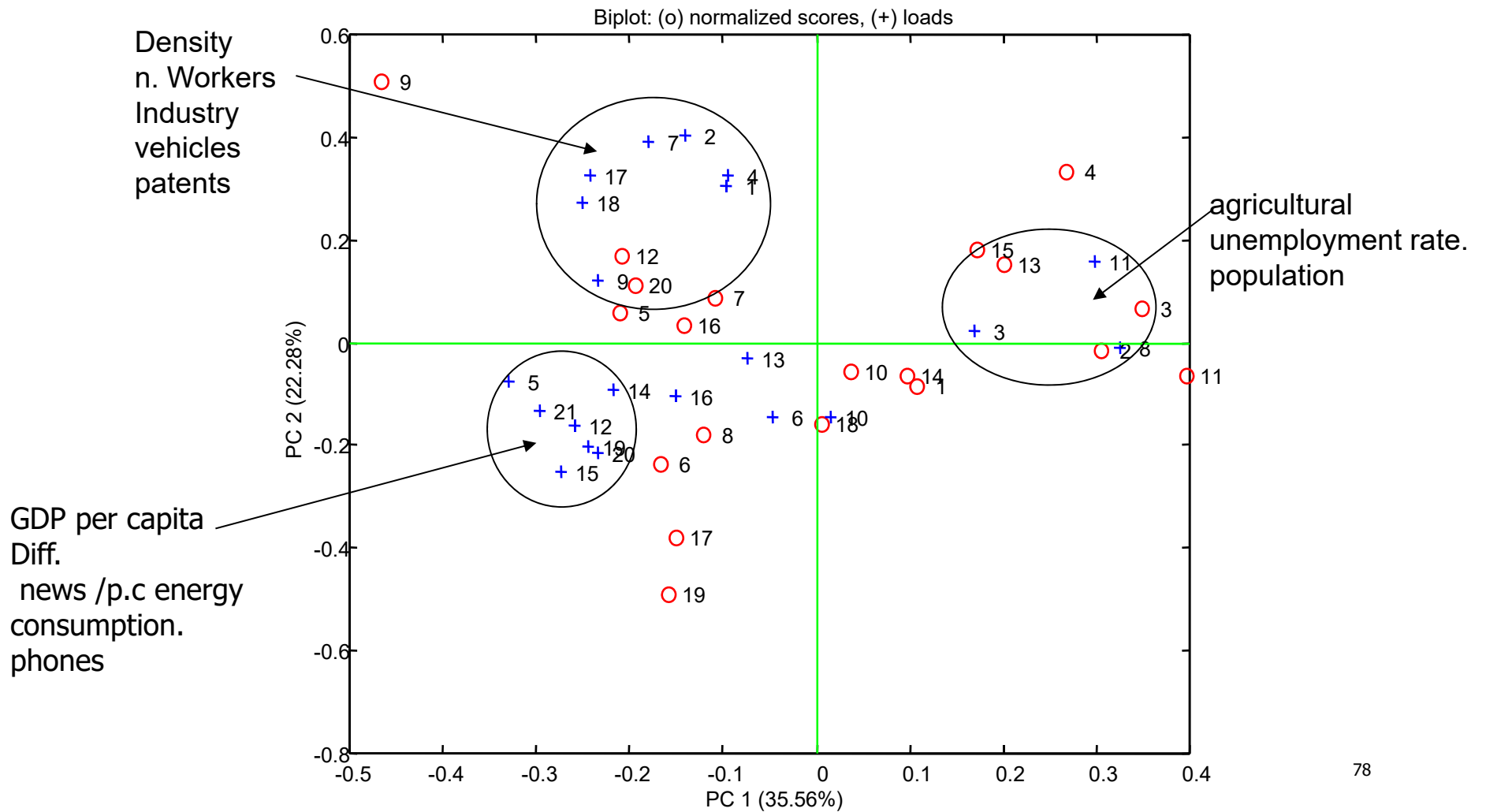


Welfare of Italian regions

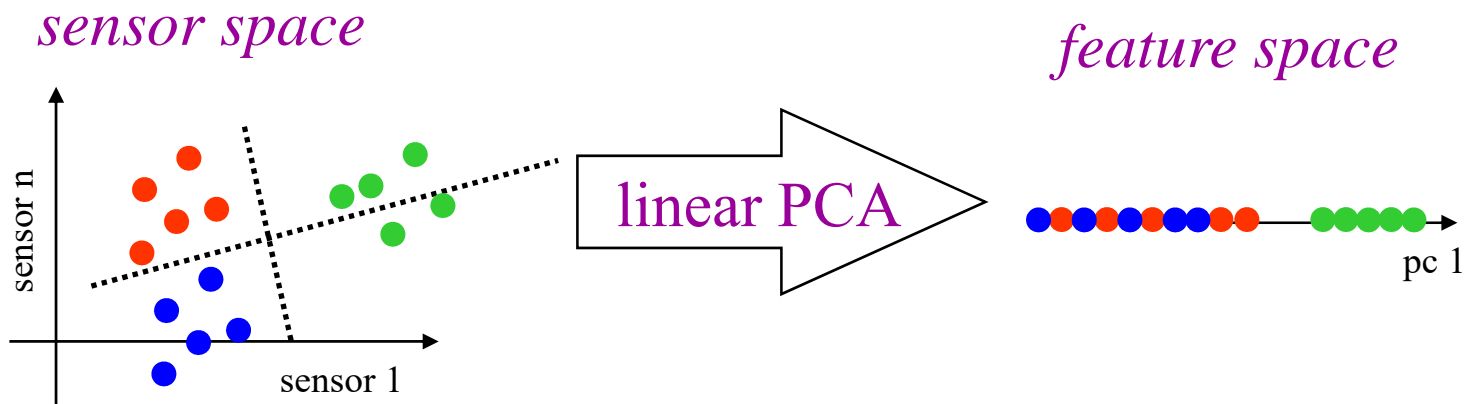
Based on 21 geographic and economics data



PCA bi-plot



PCA



PCA limitations

- The PCA representation is driven by the data characteristics in covariance matrix
- If the data are not normally distributed the covariance matrix does not satisfy the statistics of data, so the PCA representation is formally incorrect
- The score plot of the PCA is a linear projection from one to N dimension space to one dimension in the space 2 or 3. We can have false projection effects involving classification errors

Partial Least Squares (PLS)

Partial Least Squares
PLS toolbox di MATLAB

From PCR to PLS geometric approach

- The PCR solution is through the decomposition of the data matrix in the matrix of the principal components
- The principal components are the directions, in the space of the variables X , maximizing the variance and generate a base in which the X data are not correlated
- PCR in the principal components has new variables (not correlated) so becomes more easily solved.
- In PLS algorithm also the Y matrix is decomposed into principal components and principal components of X are rotated in the direction of maximum correlation to the principal components of Y
- PLS has latent variables, similar to the principal components maximizing the variance of both Y and X

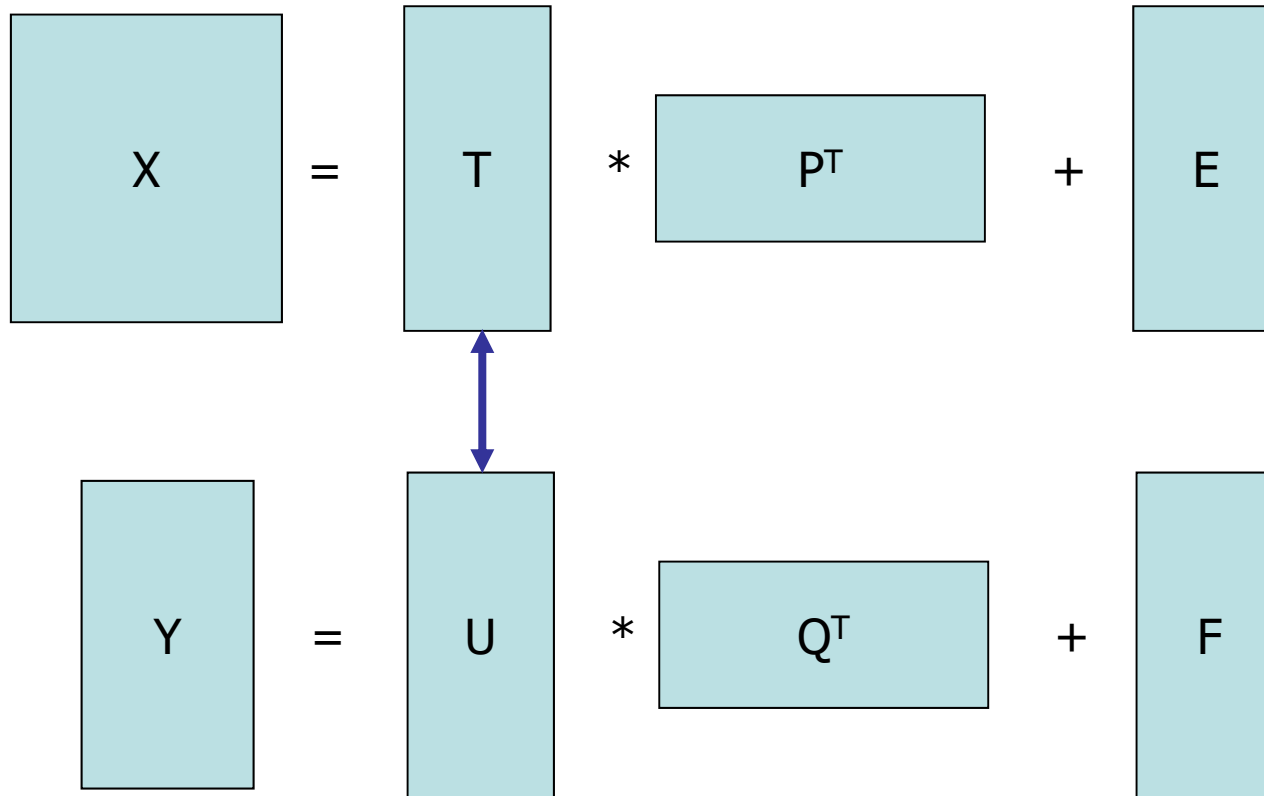
PLS importance

Partial least squares regression (PLS regression) is a statistical method that bears some relation to principal components regression; instead of finding hyperplanes of maximum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Because both the X and Y data are projected to new spaces, the PLS family of methods are known as bilinear factor models. Partial least squares Discriminant Analysis (PLS-DA) is a variant used when the Y is categorical.

PLS is used to find the fundamental relations between two matrices (X and Y), i.e. a latent variable approach to modeling the covariance structures in these two spaces. A PLS model will try to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space. PLS regression is particularly suited when the matrix of predictors has more variables than observations, and when there is multicollinearity among X values. By contrast, standard regression will fail in these cases (unless it is regularized).

The PLS algorithm is employed in partial least squares path modeling, a method of modeling a "causal" network of latent variables (causes cannot be determined without experimental or quasi-experimental methods, but one typically bases a latent variable model on the prior theoretical assumption that latent variables cause manifestations in their measured indicators).

PLS latent variables computation



The principal components are calculated maximising the correlation between T and U and their variance

$$\max [corr^2(U, T), var(U) \cdot var(T)]$$

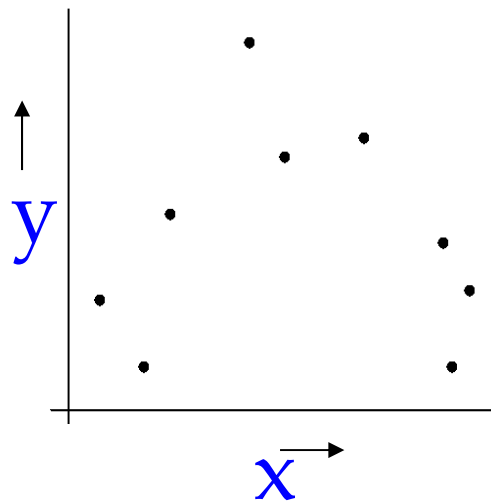
PLS Overfitting

- PLS has overfitting
- The number of latent variables must be optimized in a cross-validation process
- The overfitting is given by the fact that the latent variable k is obtained by fitting the subspace of dimension $k + 1$. The latent variables are not **orthogonal** to each other, there is no limit to the possibility of fitting the data in calibration.
- The cross-validation sets the number latent variables accuracy estimated on the validation set. Normally this value is larger than the error obtained from the model on the calibration data
- Such errors are quantified by variables:
 - RMSEC Root Mean Square Error in Calibration
 - RMSECV Root Mean Square Error of Calibration in Validation

Linear or no-linear model

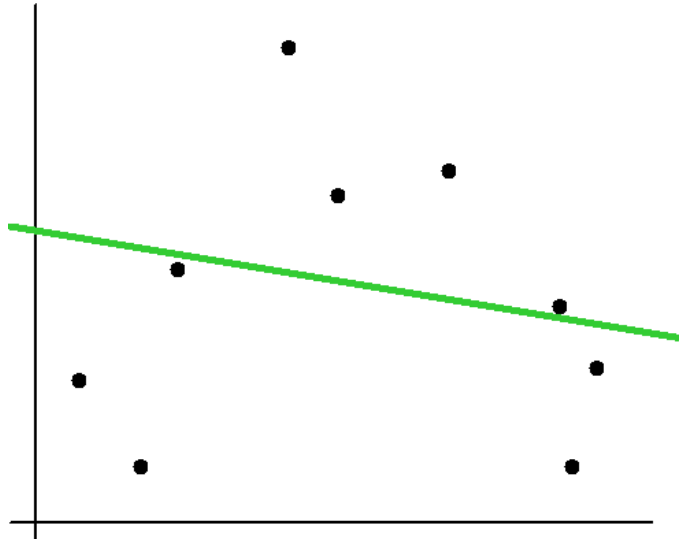
The validation problem

- Which is the best function that describes the experimental data?
- The One that allows you to predict with minimal error variables that have not been used to build the model.
- The operation that allows us to estimate this error is called cross-validation.
- Example: Consider the following information: $y=f(x)+e$
 - What is the best function that describes the relationship between x and y?

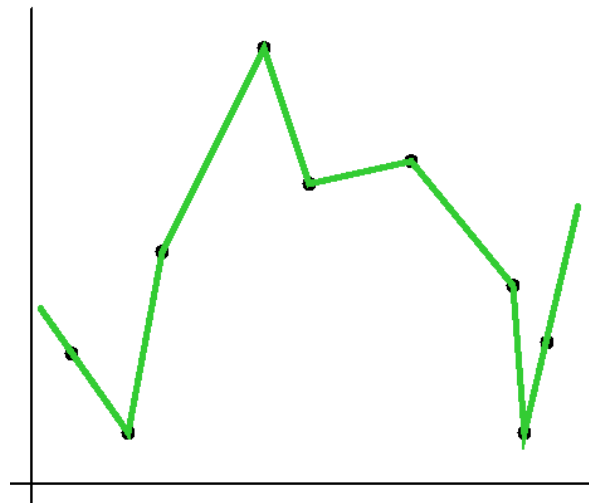
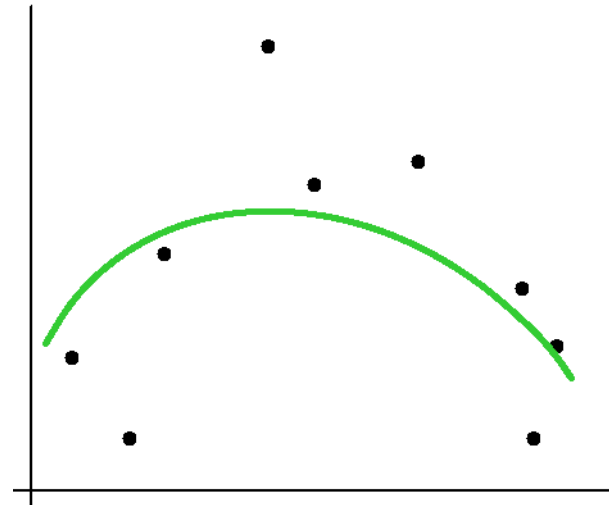


solutions

linear



No-linear



No-linear

test method

- The data set is divided into two
- The model is determined on a subset of the data (calibration with training set)
- The error is evaluated on the subset (test set)
- The prediction of the test set gives significance to the model. The data were not used for calibration. So the model can be used in the real world to estimate unknown data.

Regression predictors

- PRESS- Predicted Sum of Squares

$$PRESS = \sum_i (y_i^{LS} - y_i)^2$$

- RMSEC - Root Mean Square error of calibration

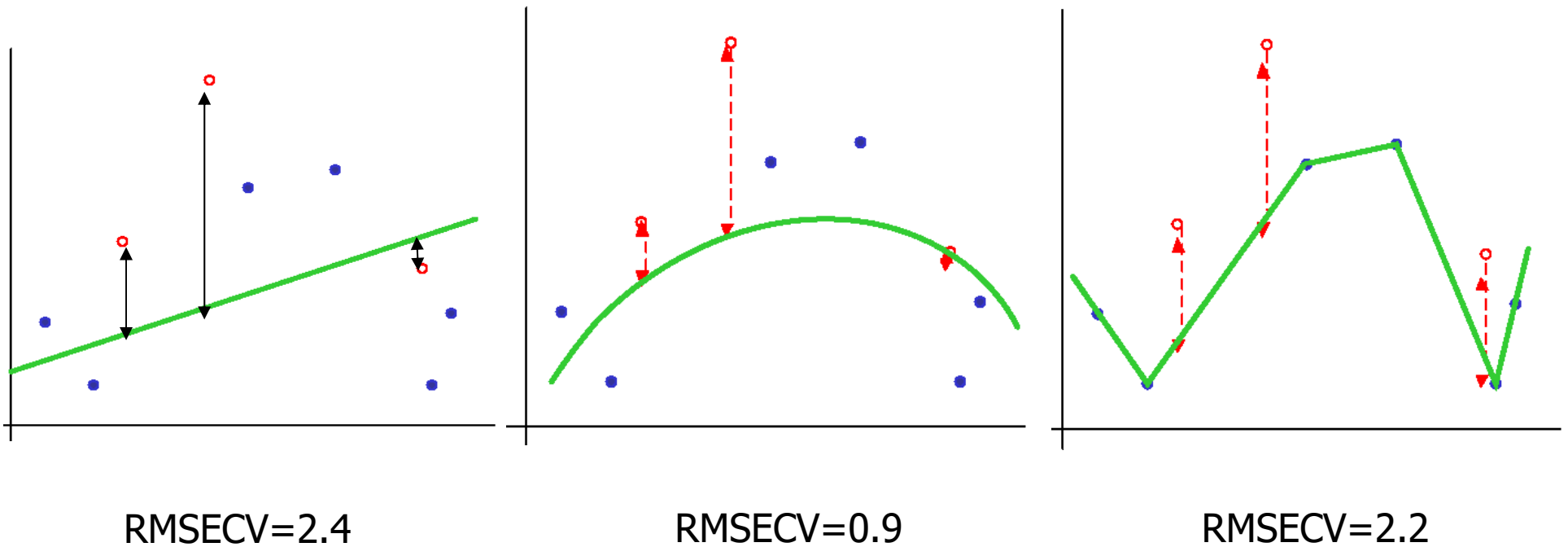
$$RMSEC = \sqrt{\frac{PRESS}{N}}$$

- RMSECV - Root Mean Square error of Cross-Validation

$$RMSECV_k = \sqrt{\frac{PRESS_k}{N}}$$

Test methods application

The data marked in red are the test set. The model is calculated on the remaining data (blue dots).
The error on the test data is evaluated as RMSECV

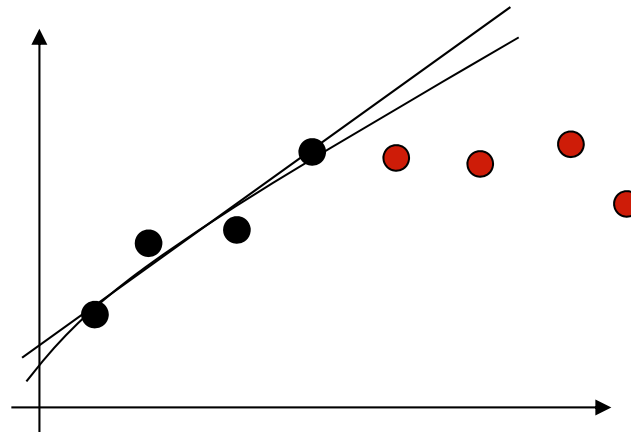


Discussion

- The best method is moderately non-linear (quadratic)
- The linear method has mistakes both in calibration and testing
- The highly non-linear method has a calibration error null but a high testing error. Such a model is "too specialized" in describing the calibration data and is not able to generalize.
- This effect is called overfitting and is typical in the case of highly non-linear models.

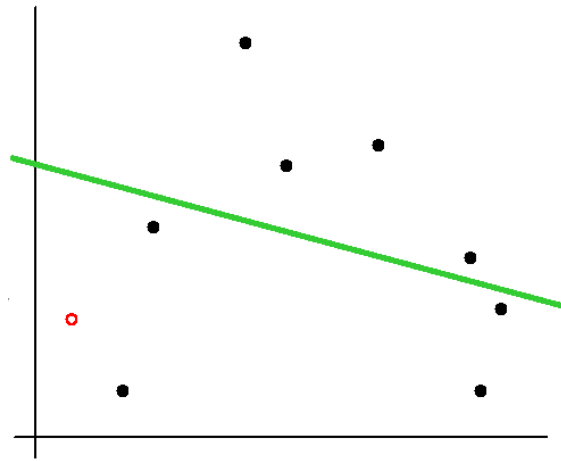
Some consideration on the test-set

- The method is very simple but requires several sets of data.
- The selection of the data is not easy in general should be done randomly but you have to avoid the two sets unbalancing
- You should check that the two sets have the same variance and the same average
- If the two sets are uncorrelated there may be apparent overfitting phenomena
- Apart from simple cases, usually the models fail in the prediction of measurements outside the range for calibration.

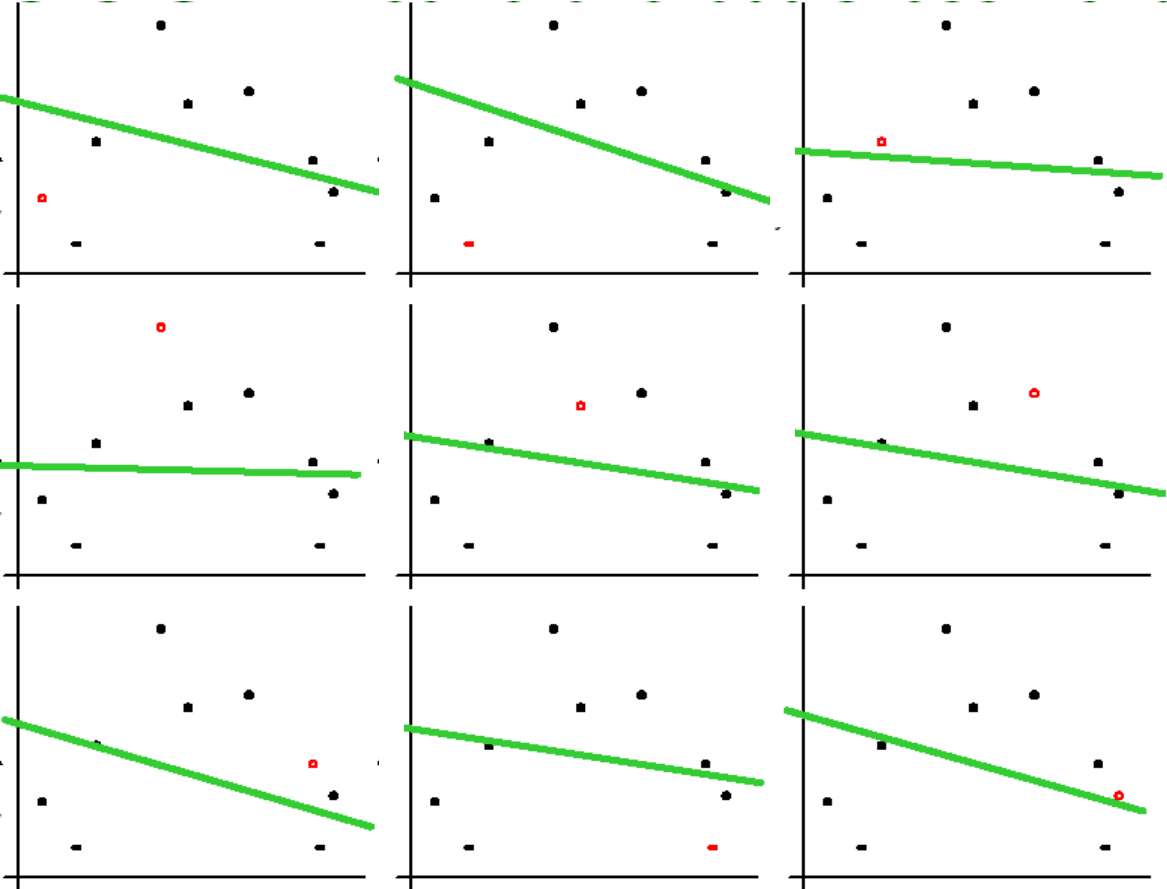


Leave-One-Out cross-validation

- When the number of data becomes small it is necessary to use other strategies for the selection of the feature and the error estimation.
- The most used method is the leave-one-out
- Leave-one-out cross-validation (LOOCV) is a particular case of leave-p-out cross-validation with $p = 1$. The process looks similar, however with cross-validation you compute a statistic on the left-out sample(s), while in the other case you compute a statistic from the kept samples only.
- LOO cross-validation does not have the problem of excessive compute time as general LpO cross-validation

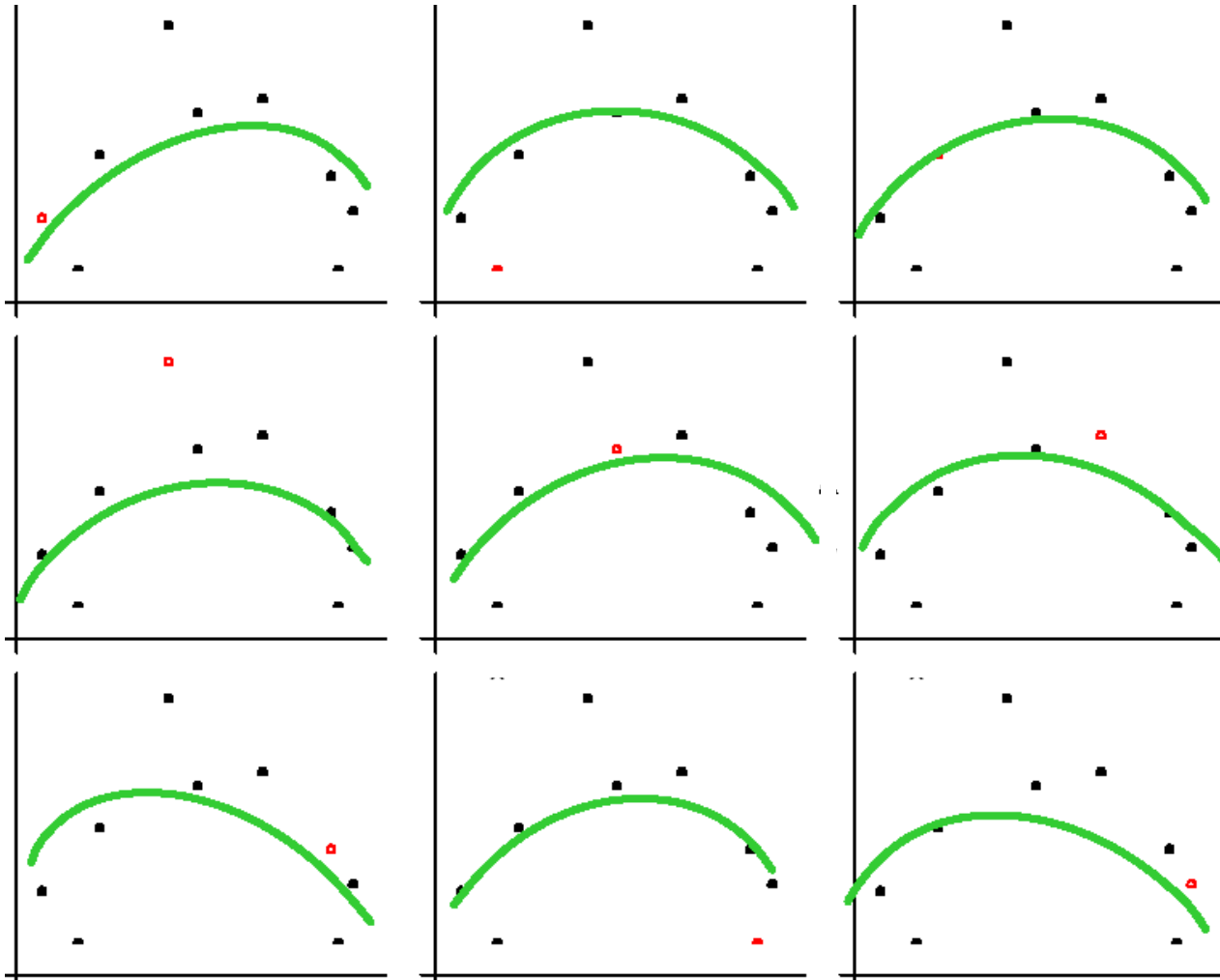


LOO linear model



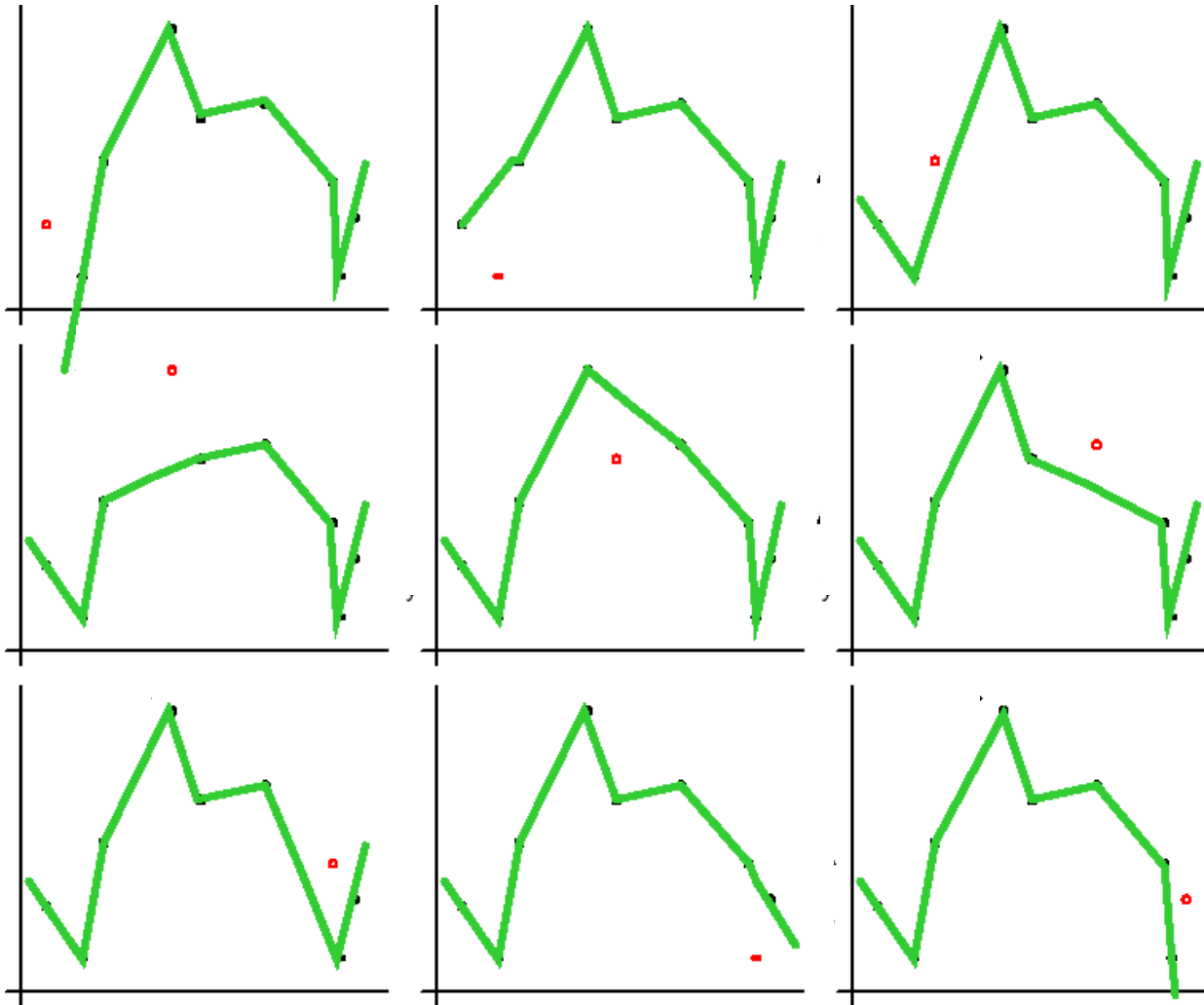
RMSECV=2.12

LOO no linear model



RMSECV=0.96

LOO highly no linear model



RMSECV=3.33

test – LOO Comparison

- LOO provides a better estimation of the prediction error than the test set whose error estimate is unreliable.
- LOO takes full advantage of the entire data set.
- Obviously LOO is the method with minimum validation.
- For sets of large dimensions LOO is expensive from the points of view of the calculation.
- It can be "softened" considering more than k data sets.

Matlab PLS toolbox *modlgui*

- data: 4 sensors TSMR for the measurement of octane and toluene

MODL_File

- Load Data
- Load Model
- Load Scale
- Load Labels
- Save Data
- Save Test
- Save Model
- Print Info
- Preferences
- Clear Data
- Clear Model
- Exit MODL

Linear Regression

File Edit View Insert Tools Window Help

MODL_File

Var: s0,nt0
Data: modeled (calibration set)
Size: 24 by 4, 24 by 2
Samp Lbls:
Var Lbls:

Model: calibrated on loaded data
Method: SIMPLS
LV(s): 3
Data: 24 by 4, 24 by 2
Scaling: autoscaled

No. LVs: 3 show parameters

calc
apply

plots

press
scores
loads
biplot
data

| Latent Variable | Percent Variance X-Block | | Percent Variance Y-Block | |
|-----------------|--------------------------|--------|--------------------------|-------|
| | This LV | Cum | This LV | Cum |
| 1 | 93.48 | 93.48 | 47.70 | 47.70 |
| 2 | 6.50 | 99.97 | 51.69 | 99.38 |
| 3 | 0.02 | 100.00 | 0.14 | 99.53 |
| 4 | 0.00 | 100.00 | 0.02 | 99.55 |

Regression Parameters

File Edit View Insert Tools Window Help

Scaling: autoscale

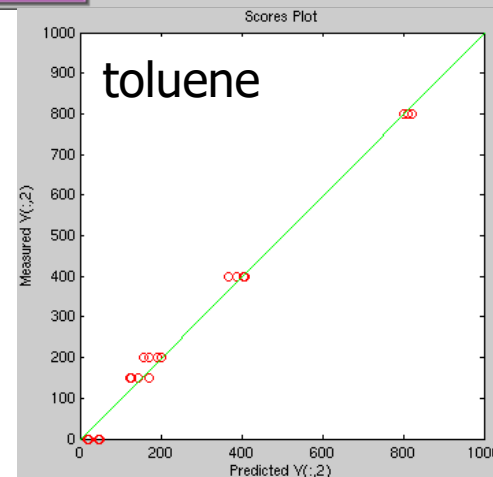
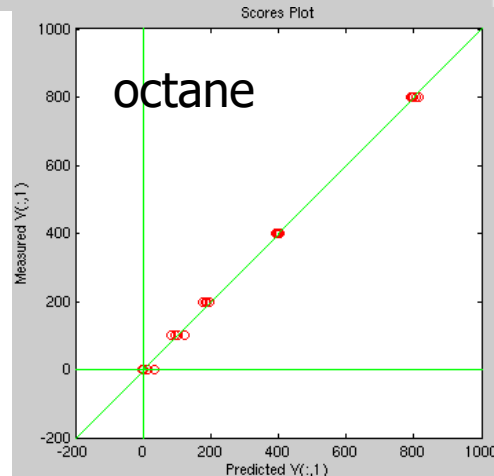
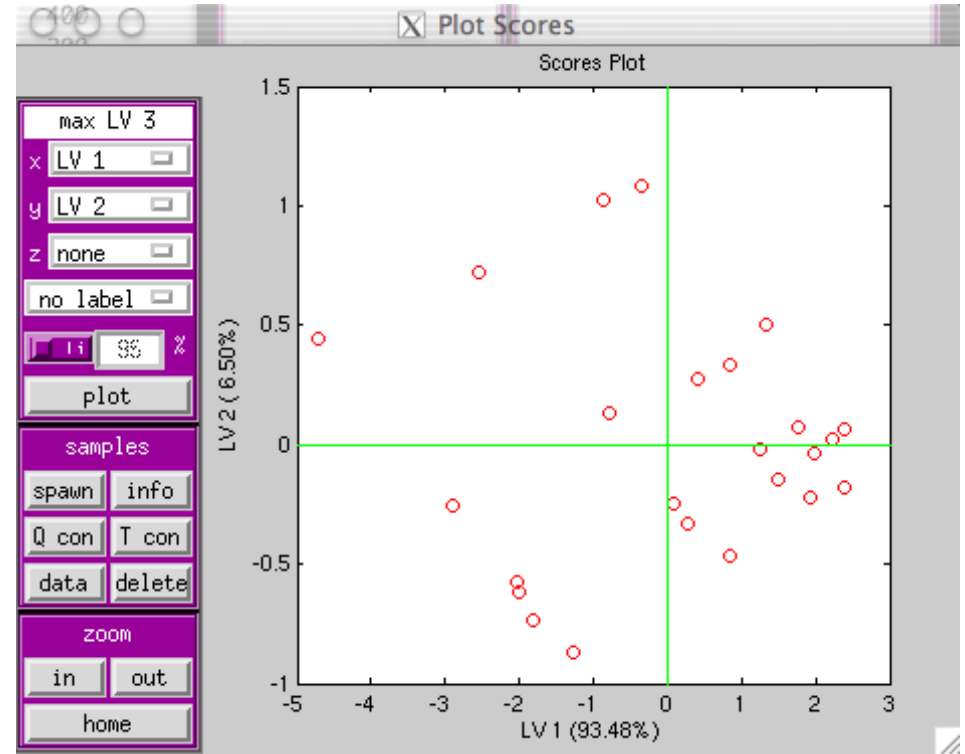
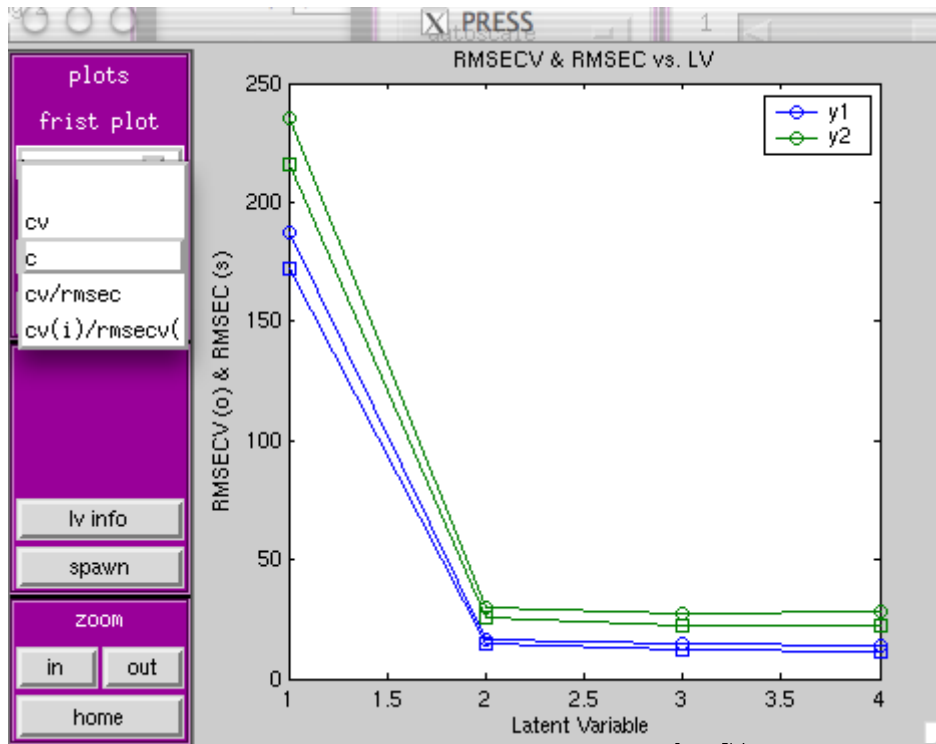
Regression: SIMPLS

Cross Validation: leave one out

Max LVs: 1 to 4

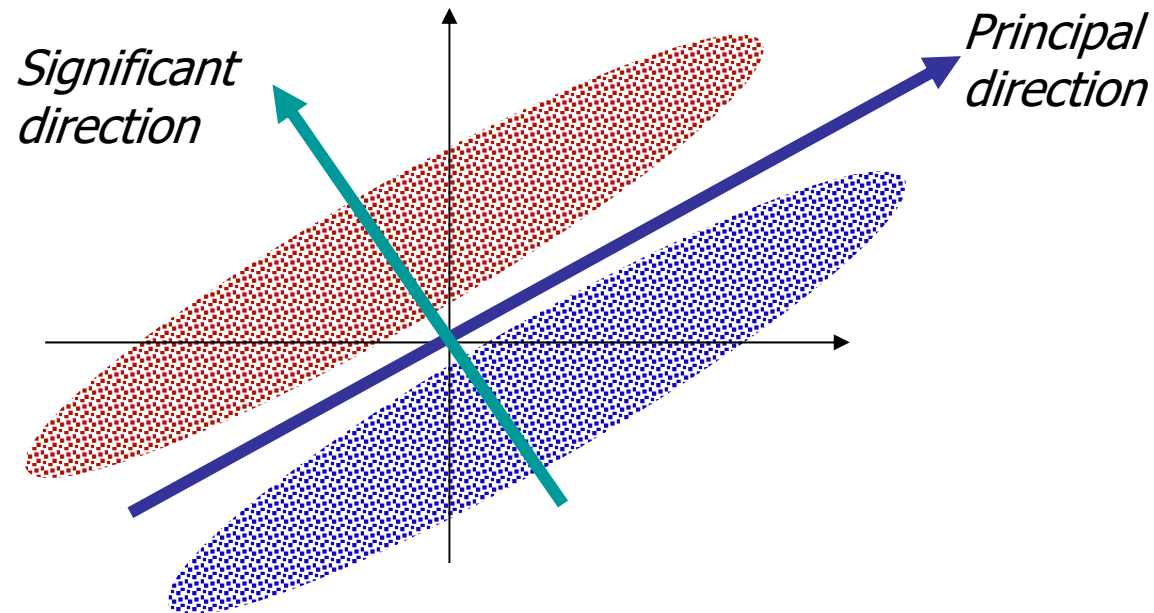
leave one out
venetian blinds
contiguous block
random subsets

modlgui



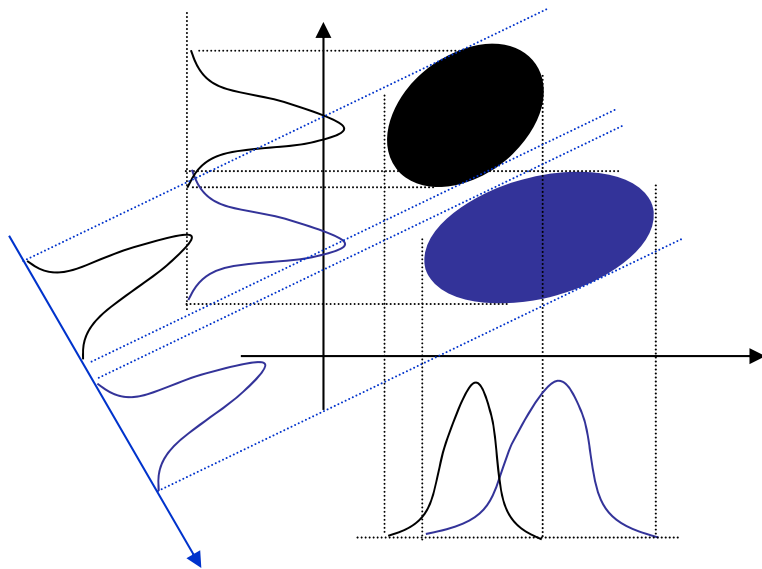
Principal components and significant directions

- The principal components are the principal axes of the ellipsoid on the covariance matrix, nothing assures the fact that these directions are important for the problem under consideration
- The "important" direction can be found using a "supervised" view ie highlighting some properties of the data set



Linear discriminant analysis (LDA)

- a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification



- There is a class of basic vectors (other than the PC) where the separation between classes is maximum
- If there are more classes you can introduce more directions
- Discriminatory directions are linear combination of real variables, you can study the contribution of each variable to the discriminant direction.

PLS-Discriminant Analysis (PLS-DA)

- PLS is the ideal tool for the solution of linear classification problems.
- Minimizing the classification error, through the score and loading plots you can study what are the patterns of the variables that mostly contribute to the classification.

PLS-DA

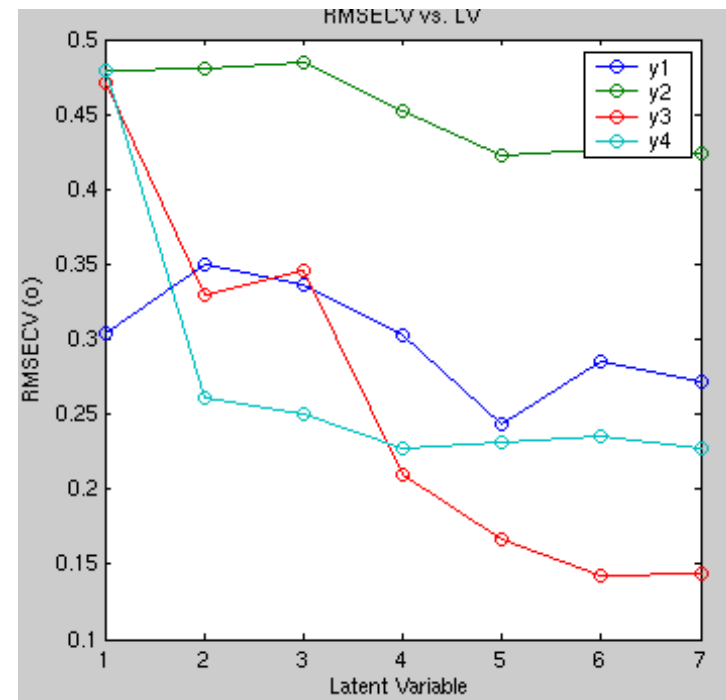
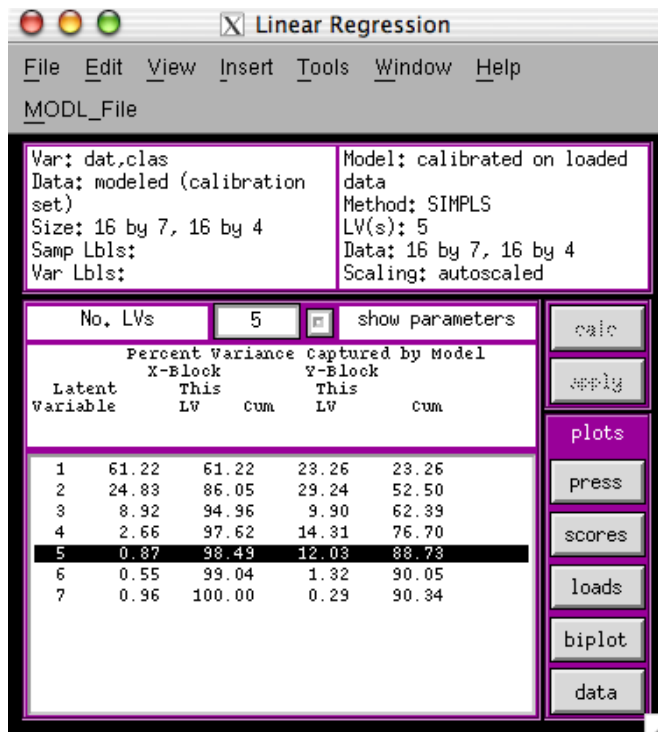
fertilizers methods for apples

- Three fertilizer methods for apples
- 1-Urea, 2-calcium nitrate and potassium, 3- ammonium sulphates 4- One control
- four classes
- Each apple is characterized by a pattern of seven features:
- Total nitrogen, seed nitrogen, phosphorus, potassium, calcium, magnesium, weight

| Y | | | | | X | | | | | | |
|---------|------|--------------------|----------------------|---|----------------|--------------|-------------|-----------|---------|-----------|--------|
| control | urea | potassium nitrates | ammonium and sulfate | | total nitrogen | pit nitrogen | phosphorous | potassium | calcium | magnesium | weight |
| 1 | 0 | 0 | 0 | 0 | 3240 | 1663 | 836 | 8747 | 218 | 388 | 97.3 |
| 1 | 0 | 0 | 0 | 0 | 3077 | 1663 | 891 | 8460 | 249 | 372 | 75.8 |
| 1 | 0 | 0 | 0 | 0 | 3205 | 1770 | 831 | 8575 | 261 | 376 | 78.5 |
| 1 | 0 | 0 | 0 | 0 | 3330 | 1755 | 889 | 8330 | 209 | 367 | 77.5 |
| 0 | 1 | 0 | 0 | 0 | 3755 | 1915 | 842 | 10375 | 145 | 408 | 108.2 |
| 0 | 1 | 0 | 0 | 0 | 5037 | 2180 | 930 | 10047 | 172 | 420 | 103.2 |
| 0 | 1 | 0 | 0 | 0 | 4753 | 2137 | 945 | 10447 | 160 | 421 | 95.3 |
| 0 | 1 | 0 | 0 | 0 | 4453 | 1967 | 850 | 9677 | 206 | 396 | 93.5 |
| 0 | 0 | 1 | 0 | 0 | 4200 | 2063 | 869 | 11190 | 184 | 398 | 111.8 |
| 0 | 0 | 1 | 0 | 0 | 5915 | 2050 | 1016 | 12060 | 184 | 461 | 109.7 |
| 0 | 0 | 1 | 0 | 0 | 5193 | 2210 | 958 | 11733 | 179 | 428 | 117.3 |
| 0 | 0 | 1 | 0 | 0 | 5347 | 2167 | 919 | 11910 | 191 | 420 | 99.6 |
| 0 | 0 | 0 | 1 | 0 | 5157 | 2357 | 1062 | 10210 | 136 | 401 | 86.7 |
| 0 | 0 | 0 | 1 | 1 | 7440 | 2975 | 1261 | 10820 | 122 | 446 | 85.5 |
| 0 | 0 | 0 | 1 | 1 | 6950 | 2527 | 1137 | 9710 | 191 | 408 | 70.8 |
| 0 | 0 | 0 | 1 | 1 | 4445 | 2075 | 846 | 8820 | 161 | 334 | 81.1 |

PLS-DA

fertilizers methods for apples



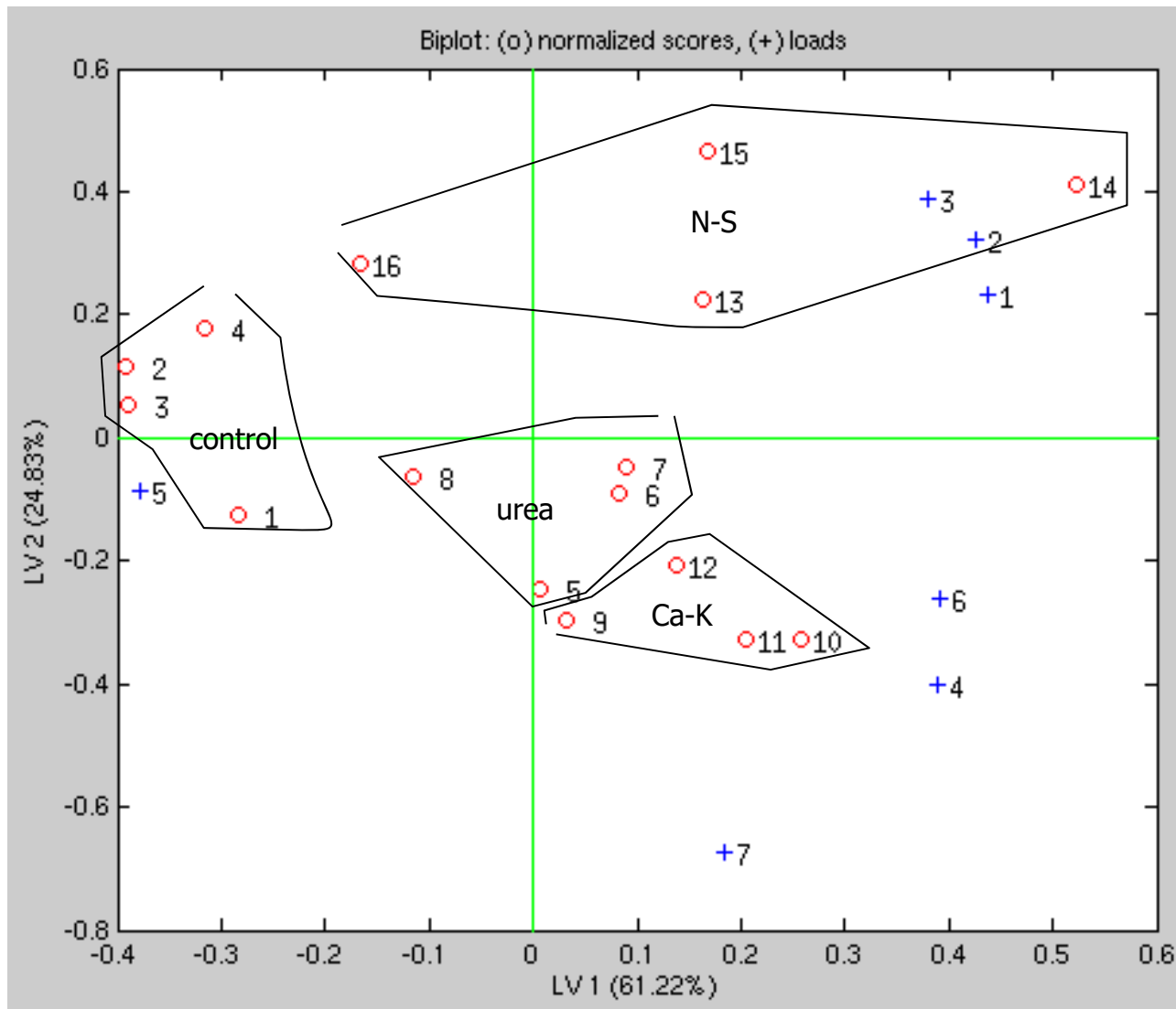
PLS-DA

fertilizers methods for apples

| Y true | | | | Y estimated | | | |
|--------|---|---|---|---------------|---------------|---------------|---------------|
| 1 | 0 | 0 | 0 | 0.7839 | 0.4456 | -0.0168 | -0.2128 |
| 1 | 0 | 0 | 0 | 1.1728 | -0.3482 | 0.1035 | 0.0718 |
| 1 | 0 | 0 | 0 | 0.8882 | 0.1623 | 0.0387 | -0.0892 |
| 1 | 0 | 0 | 0 | 0.8729 | 0.0584 | -0.2114 | 0.2801 |
| 0 | 1 | 0 | 0 | 0.0515 | 0.9332 | 0.0552 | -0.0398 |
| 0 | 1 | 0 | 0 | 0.0116 | 0.9322 | -0.0086 | 0.0648 |
| 0 | 1 | 0 | 0 | 0.0748 | 0.7485 | 0.0089 | 0.1678 |
| 0 | 1 | 0 | 0 | 0.1801 | 0.7226 | 0.0919 | 0.0053 |
| 0 | 0 | 1 | 0 | 0.0482 | -0.1203 | 0.9887 | 0.0835 |
| 0 | 0 | 1 | 0 | 0.0820 | 0.2404 | 0.8671 | -0.1895 |
| 0 | 0 | 1 | 0 | -0.0390 | -0.0669 | 1.0746 | 0.0313 |
| 0 | 0 | 1 | 0 | -0.1771 | 0.0942 | 0.9599 | 0.1230 |
| 0 | 0 | 0 | 1 | 0.1673 | -0.0846 | 0.1036 | 0.8137 |
| 0 | 0 | 0 | 1 | -0.2136 | 0.1390 | -0.0245 | 1.0990 |
| 0 | 0 | 0 | 1 | 0.1372 | -0.0736 | 0.0300 | 0.9064 |
| 0 | 0 | 0 | 1 | -0.0410 | 0.2174 | -0.0608 | 0.8844 |

PLS-DA

fertilizers methods for apples



The Scores show:

- The separation between the four groups
- From the control we have two directions: N-S and urea//Ca-K

The loadings show:

- The N-S treatment increases the total nitrogen and phosphorus in the seed
- The treatments with urea and Ca-K increase potassium, magnesium, and the weight of the fruit
- The greater amount of calcium is found in the control apples