

Session 4.1

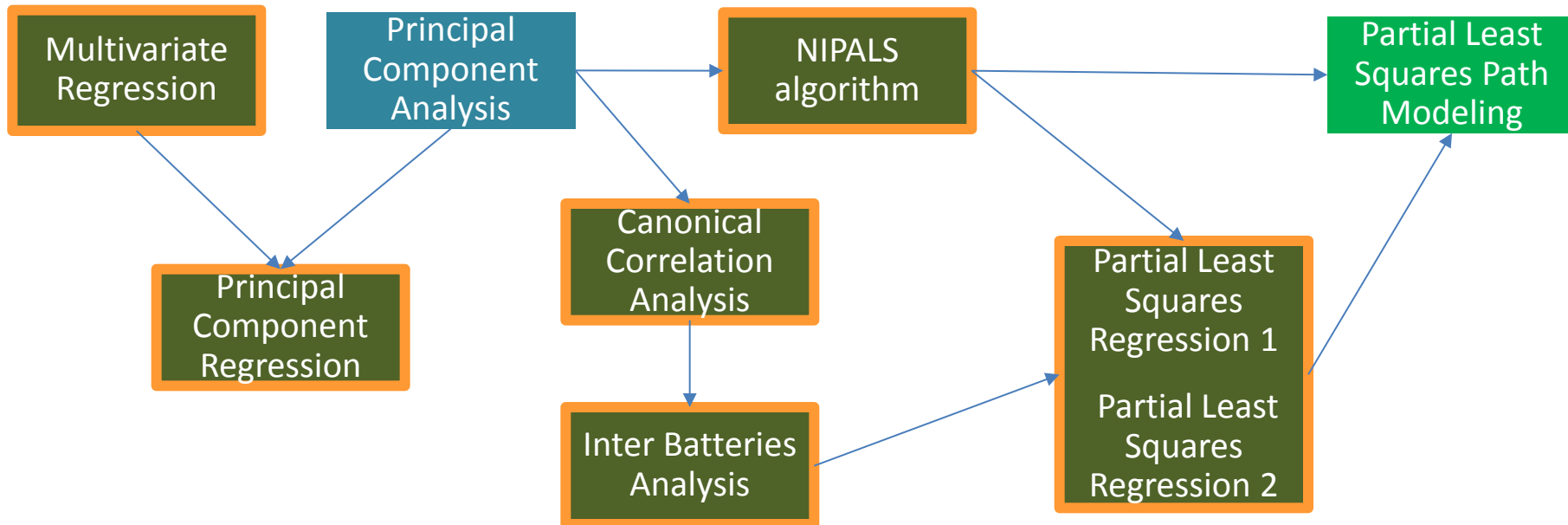
NIPALS

Course on Multivariate Modeling

Tomàs Aluja-Banet

tomas.aluja@upc.edu

The plan of the course



Multivariate Descriptive techniques

Criterion to optimize:

- in PCA, component with max. variance $\max \text{var}(t_h)$
- in CCA, components with max. correlation
- in ...

Restrictions:

- on $\mathbb{R}^p \rightarrow$ normalized weights $\|w_h\| = 1$
- on $\mathbb{R}^n \rightarrow$ normalized components $\|t_h\| = 1$

NIPALS

Nonlinear estimation by Iterative Partial Least Squares, Herman Wold 1966



NIPALS is an iterative algorithm, proposed by Herman Wold, to compute the principal components.

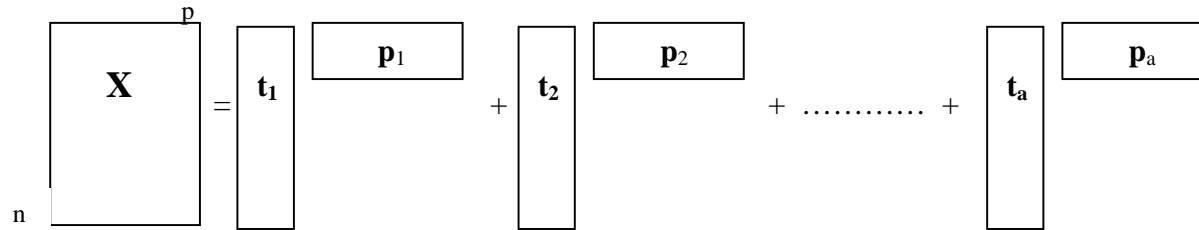
- It is based in the PCA model (=biplot).
- It allows the presence of missing values.
- Cross validation to select the number of components.
- Identification of outliers
- It is based in the singular value decomposition for computing the eigenvectors and eigenvalues of a matrix X .

Let be a matrix X of n individuals and p **centered** variables. (for the moment, without missing data)

Let be t_h the corresponding *principal components* of X and p_h the eigenvectors of R^p

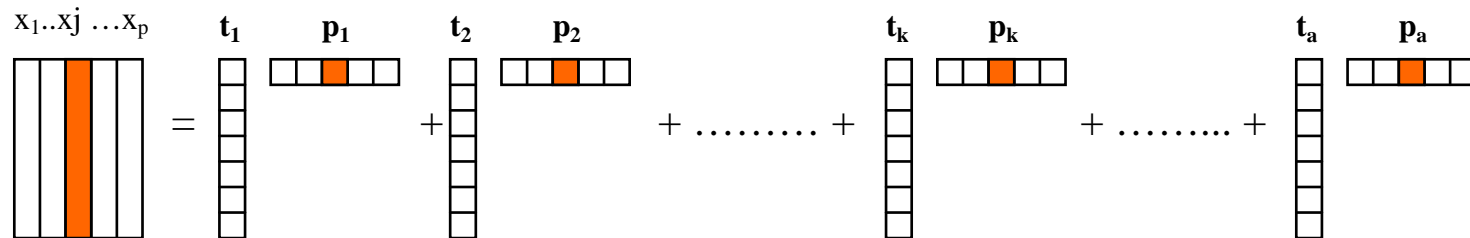
$$X = TP' \qquad X = \sum_{h=1}^a t_h p_h' \qquad \text{Let } a = \text{rang}(X)$$

PCA model: multivariate regression of X columns on t_h



$$\begin{matrix} p \\ \mathbf{X} \\ n \end{matrix} = \begin{matrix} t_1 \\ \mathbf{p}_1 \end{matrix} + \begin{matrix} t_2 \\ \mathbf{p}_2 \end{matrix} + \dots + \begin{matrix} t_a \\ \mathbf{p}_a \end{matrix}$$

Every variable can be written as linear composite as regression function (without intercept) of principal components

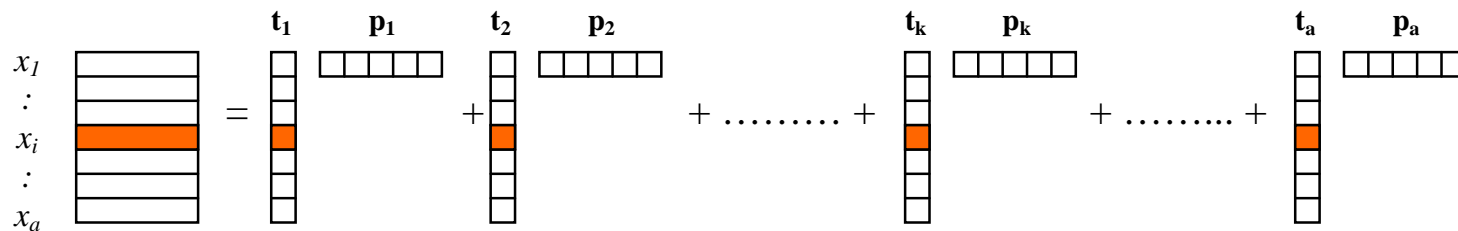


$$\begin{matrix} x_1 \dots x_j \dots x_p \\ \vdots \\ \vdots \end{matrix} = \begin{matrix} t_1 \\ \vdots \\ \vdots \end{matrix} \begin{matrix} p_1 \\ \vdots \\ \vdots \end{matrix} + \begin{matrix} t_2 \\ \vdots \\ \vdots \end{matrix} \begin{matrix} p_2 \\ \vdots \\ \vdots \end{matrix} + \dots + \begin{matrix} t_k \\ \vdots \\ \vdots \end{matrix} \begin{matrix} p_k \\ \vdots \\ \vdots \end{matrix} + \dots + \begin{matrix} t_a \\ \vdots \\ \vdots \end{matrix} \begin{matrix} p_a \\ \vdots \\ \vdots \end{matrix}$$

$$x_j = p_{1j}t_1 + p_{2j}t_2 + \dots + p_{aj}t_a$$

PCA model: multivariate regression of X rows on p_h

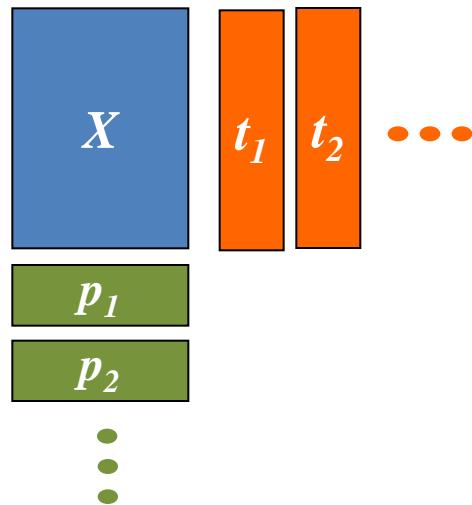
Every row can be written as linear composite, regression function (without intercept) of eigenvectors



$$x_i = t_{1i}p_1 + t_{2i}p_2 + \dots + t_{ai}p_a$$

NIPALS algorithm exploits these equivalences

NIPALS algorithm



$$X_0 = X$$

for $h = 1, 2, \dots, a$

$$t_1 = \text{rowMeans}(X_{h-1})$$

iterate till convergence of p_h

$$p_h = X'_{h-1} t_h$$

$$\|p_h\| = 1$$

$$t_h = X_{h-1} p_h$$

$$X_h = X_{h-1} - t_h p_h'$$

deflation

X centered

X complete data



In convergence

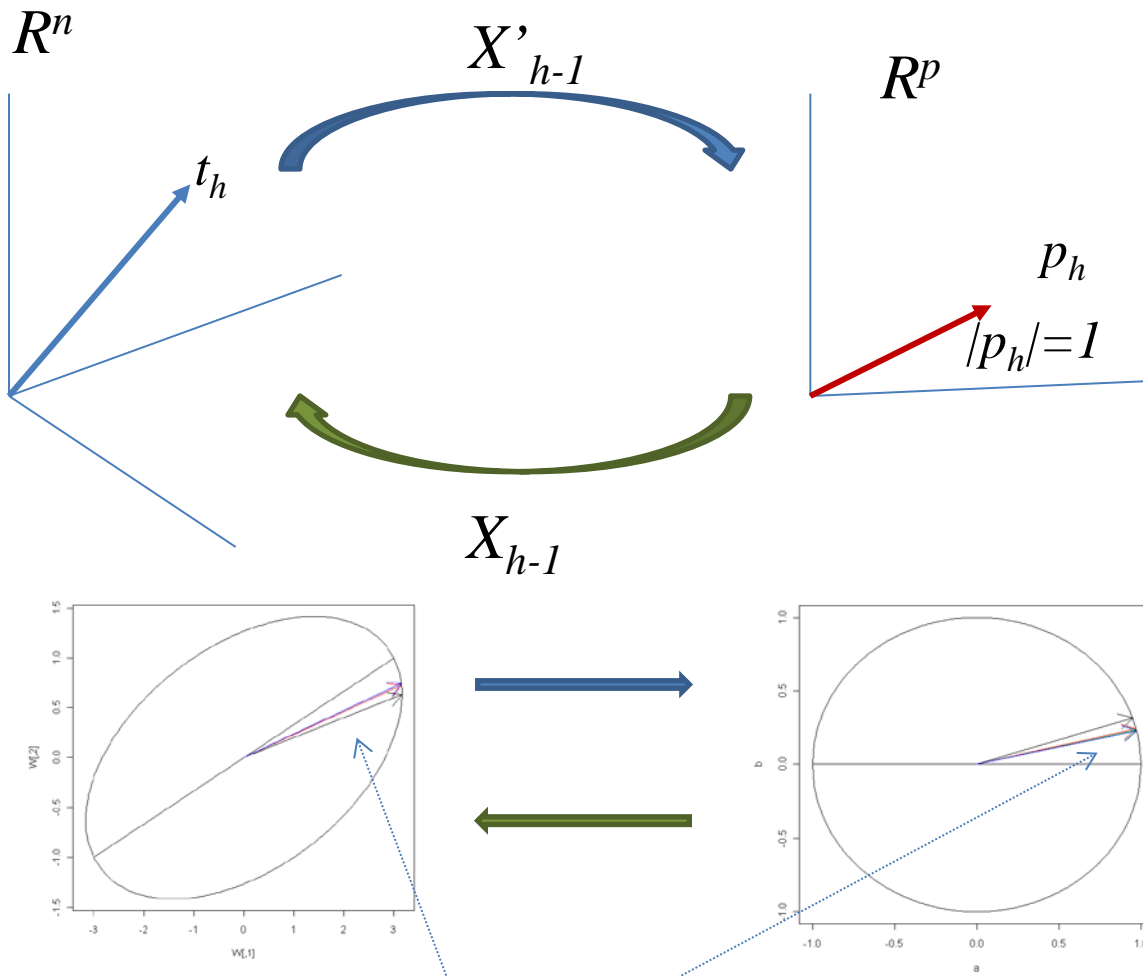
$$X'_{h-1} X_{h-1} p_h \propto p_h \quad p_h' p_h = 1$$

$$t_h = X_{h-1} p_h \rightarrow \text{Max } t_h' t_h = \lambda_h$$

p_h unitary eigenvector of $X'X$

t_h principal component

The geometry of NIPALS



In convergence

```
> X
      [,1] [,2]
[1,]      3      1
[2,]      1     -1
> t = rowMeans(X)
> for (i in 1:100) {
  p=t(X)%*%t;
  p=p/sqrt(sum(p*p));
  t=X%*%p;
  arrows(0,0,t[1],t[2],col=i)}
> print(t)
      [,1]
[1,] 3.149500
[2,] 0.743496
> sqrt(sum(t*t))
[1] 3.236068
> print(p)
      [,1]
[1,] 0.9732490
[2,] 0.2297529
> eigen(t(X)%*%X)
$values
sqrt: [1] 3.236068 1.236068
$vectors
      [,1]      [,2]
[1,] -0.9732490  0.2297529
[2,] -0.2297529 -0.9732490
```

How to deal with missing values

The idea is to transform a scalar product to a series of regressions

$$X_0 = X$$

for $h = 1, 2, \dots, a$

$$t_1 = \text{rowMeans}(X_{h-1})$$

iterate till convergence of p_h

$$p_h = X'_{h-1} t_h / t'_h t_h$$

$$\|p_h\| = 1$$

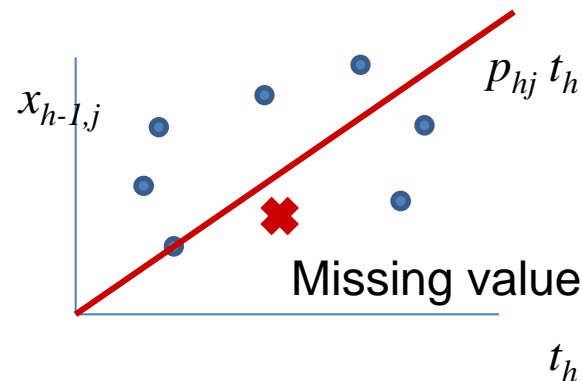
$$t_h = X_{h-1} p_h / p'_h p_h$$

$$X_h = X_{h-1} - t_h p'_h$$

Writing $p_h = X'_{h-1} t_h / t'_h t_h$

$$p_{hj} = (t'_h t_h)^{-1} t'_h x_{h-1,j} \quad j = 1, \dots, p$$

we compute the p components of vector p_h as a simple regression of each x_j on t_h component (regression without intercept)



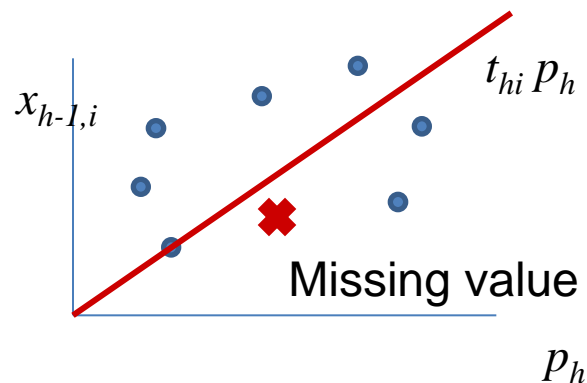
The missing values are simply discarded from the regression calculation

Dealing with missing values (2)

Likewise
$$t_h = X_{h-1} p_h / p_h' p_h \quad t_{hi} = (p_h' p_h)^{-1} p_h' x_{h-1,i} \quad i = 1, \dots, n$$

Each component of the t_h vector is a regression coefficient (without intercept) of each row of X_{h-1} on p_h eigenvector.

Missing values gives raise to absent points in the calculation of the coefficient t_{hi}



Number of components

By crossvalidation (usually by LOO):

Position the omitted X^i points as supplementary in the space of t_h components

$$T_h^{-k} = X^{-k} P_h$$

Reconstitution formula $\hat{X}_h^{-k} = T_h^{-k} P_h'$

Predicted Residual Sum of Squares

$$PRESS_h = \sum_{i,j} (x_{ij}^{-k} - \hat{x}_{h,ij}^{-k})^2$$

$$R_{cv,h}^2 = 1 - \frac{PRESS_h}{\sum_{i,j} (x_{ij} - \bar{x}_j)^2}$$

Component t_h is kept if its predicted error is decreasing (or R_{cv}^2 is increasing)

Identification of outliers

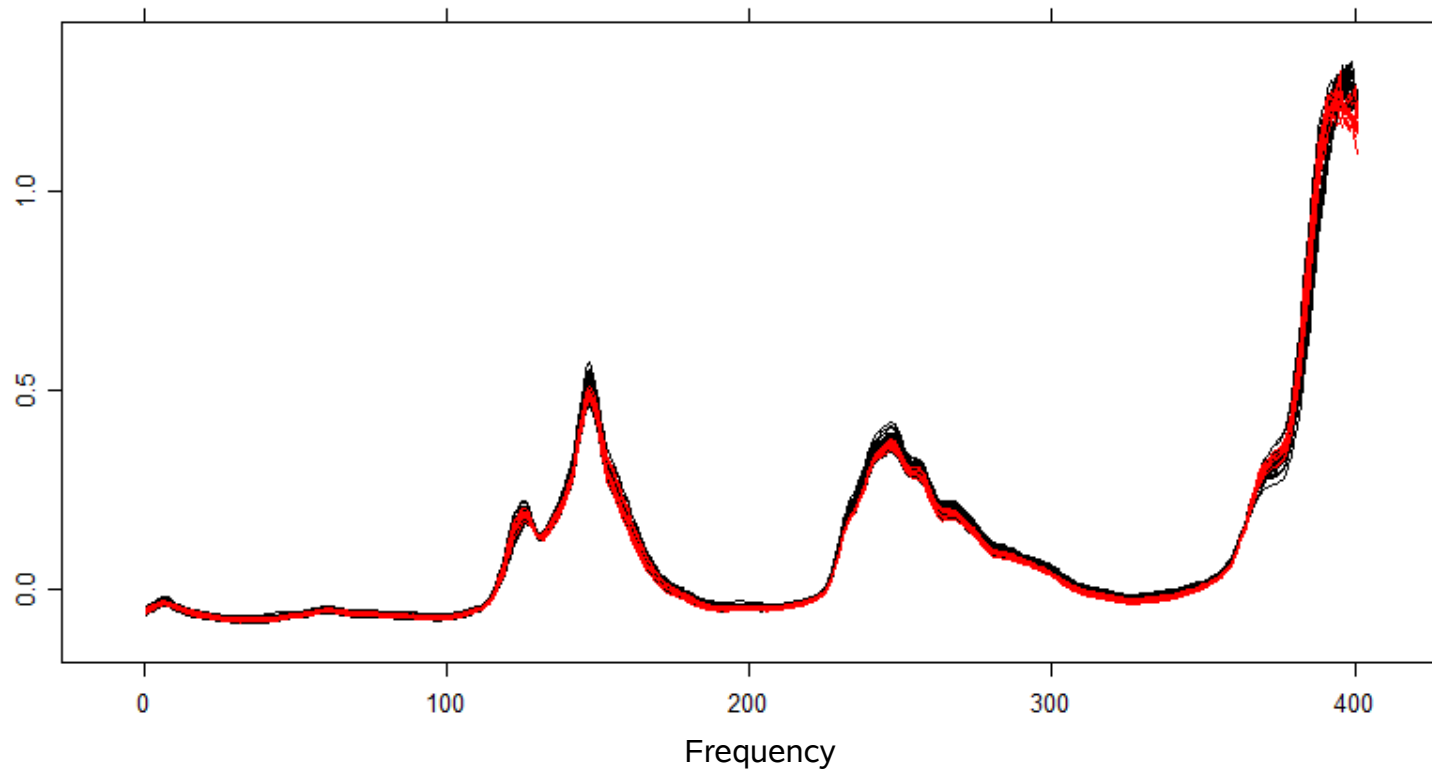
We compute the distance of every individual x_i (row of X) to its prediction obtained with the selected number of components. We relativize it respect to the average distance. It is conjectured that this ratio follows a F distribution.

$$DModX_h = \sqrt{\frac{d^2(x_i, \hat{x}_{i.h})}{\frac{1}{n} \sum_i d^2(x_i, \hat{x}_{i.h})}} \quad \approx \text{standardized residual}$$

Gasoline NIR (Near Infra Red) spectrum data

- We want to predict the Octane number of a gasoline from the NIR (Near Infra Red spectrum) of gasolines.
- The higher the octane number, the less likely is the fuel to ignite prematurely in the engine's cycle and cause the engine damage.
- Measuring the Octane number: The most common type of octane rating worldwide is the Research Octane Number (RON). RON is determined by running the fuel in a test engine with a variable compression ratio under controlled conditions, and comparing the results with those for mixtures of iso-octane and n-heptane.
- Infrared (IR) light is electromagnetic radiation with a wavelength longer than that of visible light, measured from the nominal edge of visible red light at 0.74 micrometers, and extending conventionally to 300 micrometres. Microscopically, IR light is typically emitted or absorbed by molecules when they change their rotational-vibrational movements. The infrared portion of the electromagnetic spectrum is usually divided into three regions; the near-, mid- and far- infrared, named for their relation to the visible spectrum.
- Spectroscopy: Is a technique which can be used to identify molecules by analysis of their constituent bonds. Each chemical bond in a molecule vibrates at a frequency which is characteristic of that bond. A group of atoms in a molecule may have multiple modes of oscillation caused by the stretching and bending motions of the group as a whole. The vibrational frequencies of most molecules correspond to the frequencies of infrared light. Typically, the technique is used to study organic compounds. This can be used to gain information about the sample composition in terms of chemical groups present and also its purity.

NIR spectrum

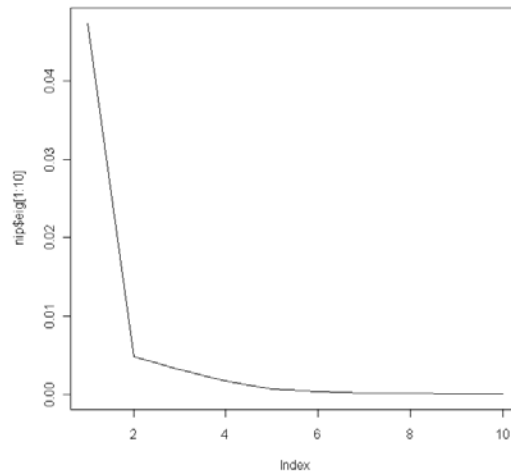


We have 60 gasolines, from which we have measured their octane number and their NIR spectrum (900nm – 1700nm), having 401 frequencies per each gasoline

We will use the first 50 gasolines as training sample, and the last 10 as test (holdout sample) (in red)

Exploring the NIR data

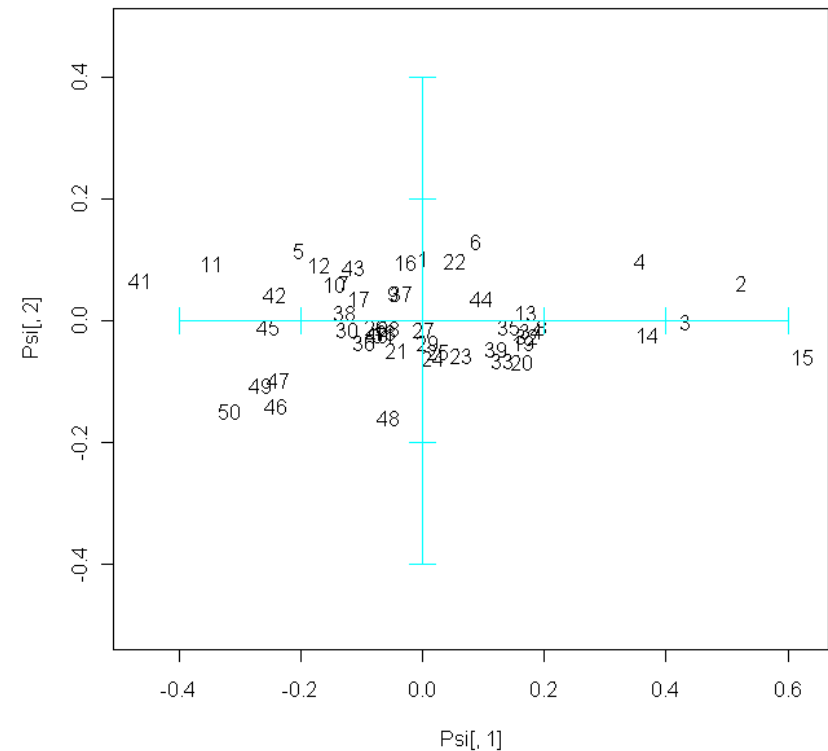
Screeplot



Total inertia: 0.05929666

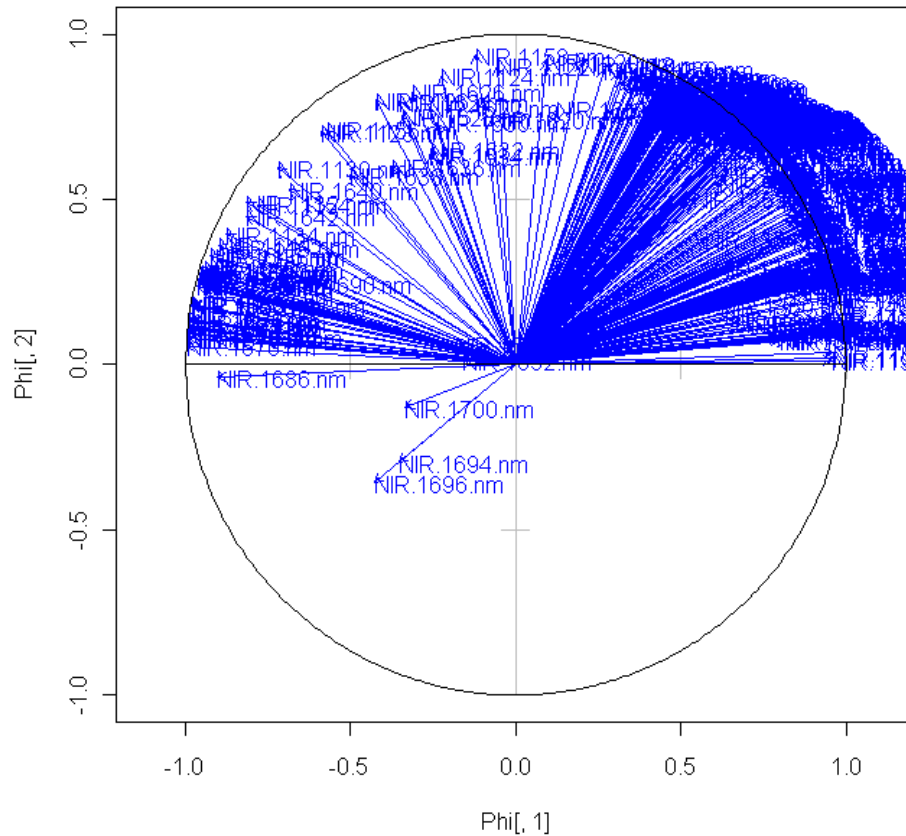
	eigenvalue	% explained	% cumulated
1	4.735e-02	7.986e+01	79.86
2	4.900e-03	8.264e+00	88.12
3	3.212e-03	5.417e+00	93.54
4	1.781e-03	3.003e+00	96.54
5	7.094e-04	1.196e+00	97.74
6	3.794e-04	6.398e-01	98.38
7	2.189e-04	3.692e-01	98.75
8	1.855e-04	3.128e-01	99.06
9	1.287e-04	2.171e-01	99.28
10	8.408e-05	1.418e-01	99.42

Plot of individuals

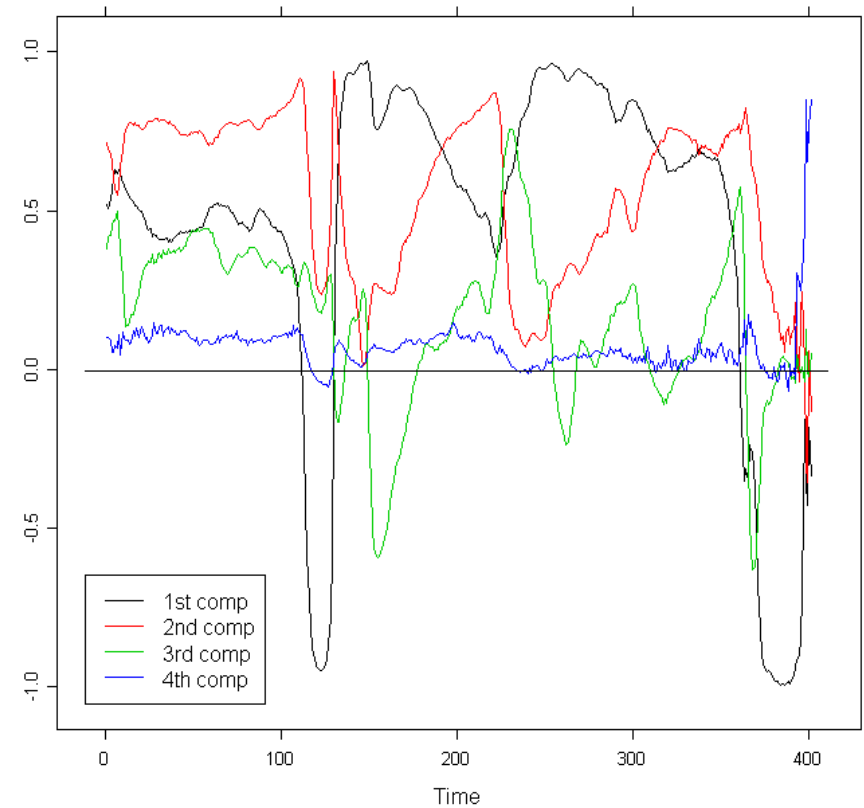


Exploring the NIR data

Plot of correlations of variables



Correlations of NIR with Nipals components



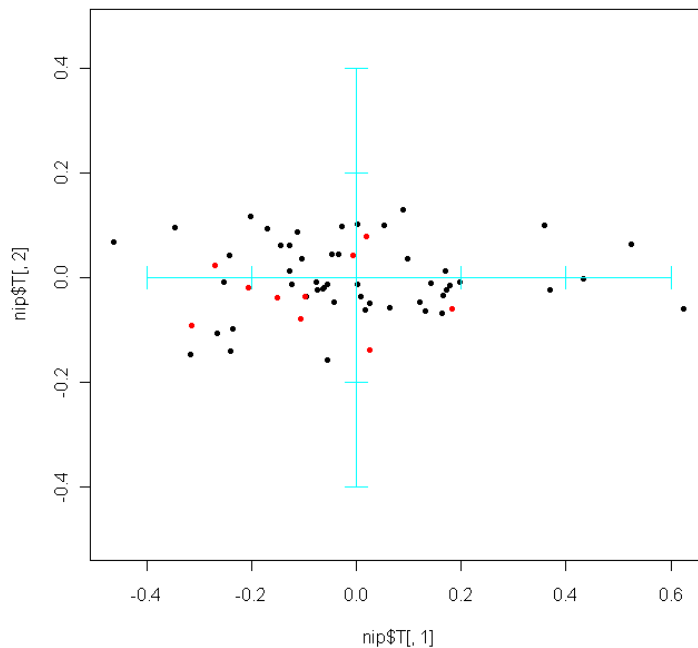
Grafting the Test data



Test individuals projected as supplementary points

Test variables projected as correlations with the projections of test individuals

Train and Test individuals



Plot of octane en train and test data

