Seconda Università di Napoli, 2015/06/25

From Principal Component Analysis to Partial Least-Squares regression

Jean-François Durand

Montpellier, France

Package: Boosted Partial Least-Squares regression

http://www.jf-durand-pls.com

The data

- p predictors (continuous or categorical) $x^1, \ldots, x^p$
- q responses (continuous or categorical) $y^1, \ldots, y^q$
  - continuous : regression model
  - q indicator variables : classification model
- n observations

Two data matrices : X n x p,
Y n x q

$$r = rank(X) \leq \min(n, p)$$

All columns (variables) are standardized (weights 1/n)

Latent variables from the X data

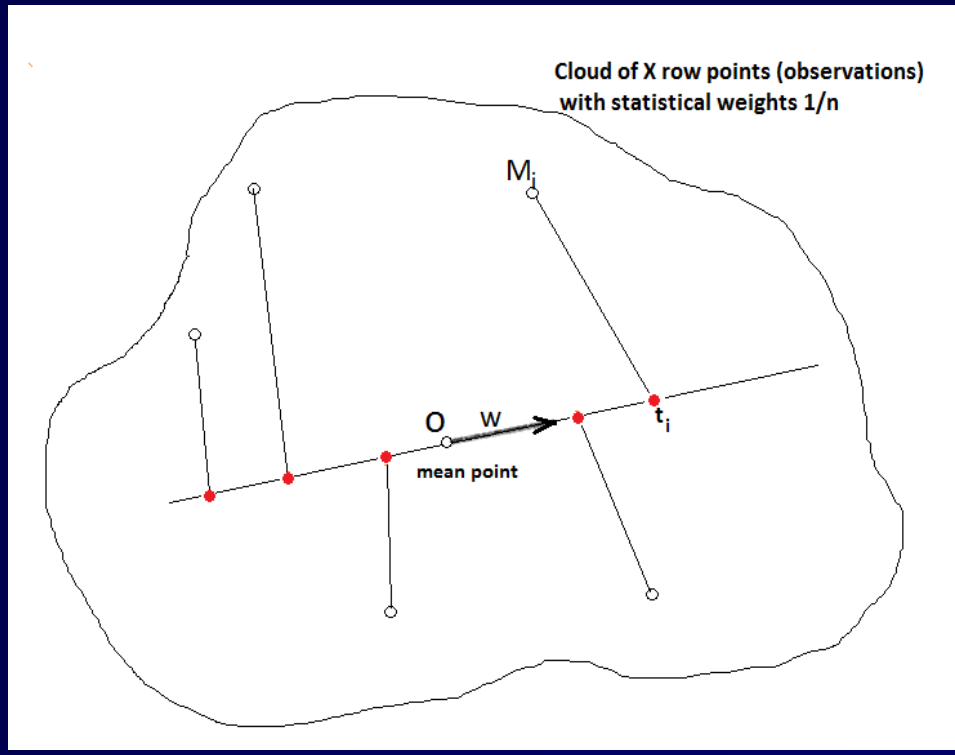
Dual interpretation of $t = Xw$

1) a latent variable: linear combination of the variables (X columns)

$t = Xw = x^1 w_1 + \ldots + x^p w_p$. The vector of weights: w

2) a component: coordinates of the projections of the n observations on the 'axis' whose direction is defined by w .

var(t) is also the inertia of the projected points, a measure of the distance from the mean point (the origin of the space of observations).



One small example to test the methods

The orange juice data

p=10 predictors (continuous)

COND (conductivity)

8 mineralogical characters

SiO2, Na, K, Ca, Mg, Cl, SO4, HCO3

and their Sum (r = 9)

q=1 response (continuous)

Heavy (a sensory descriptor)

n=24 orange juices

$$X = [x^1 x^2 x^3 x^4 x^5 x^6 x^7 x^8 x^9 x^{10}]$$ **Y**

	COND	SiO2	Na	K	Ca	Mg	Cl	SO4	HCO3	Sum	Heavy
A	376.0	17.29	19.20	1.43	44.80	26.00	29.20	18.20	199.12	355.24	9.81
B	369.6	13.39	7.52	1.43	83.20	2.34	12.40	6.50	246.96	373.74	10.68
C	168.8	42.90	8.32	4.94	24.00	4.55	2.08	54.60	63.44	204.83	7.11
D	121.6	2.34	5.60	0.78	17.20	0.00	30.40	0.00	13.68	70.00	7.05
E	472.8	24.70	36.80	3.90	58.40	25.48	25.60	60.45	261.60	496.93	12.96
F	40.8	9.75	2.56	0.39	4.00	2.21	3.12	5.20	14.64	41.87	5.70
G	234.4	26.00	20.16	1.69	21.12	11.57	34.16	26.00	80.00	220.70	9.12
H	280.0	11.57	26.40	1.30	19.20	8.97	49.60	24.70	468.48	610.22	8.73
I	200.0	28.60	10.40	1.30	32.80	10.01	8.80	18.20	1079.44	1189.55	8.43
J	168.0	25.35	7.76	1.95	24.88	5.98	4.80	13.00	89.76	173.48	7.26
K	1618.4	10.40	6.96	4.29	388.80	110.50	6.40	1534.00	321.12	2382.47	19.53
L	398.4	18.85	4.80	1.30	64.00	33.54	3.52	16.25	297.68	439.94	10.86
M	1788.0	11.96	12.80	4.94	432.00	146.90	11.52	1820.00	326.00	2766.12	19.92
N	234.4	7.54	1.92	0.52	56.80	2.08	2.88	20.80	161.04	253.58	8.76
O	804.0	10.40	2.96	2.47	160.00	50.70	7.28	377.00	325.04	935.85	16.53
P	1823.2	16.90	6.80	3.25	522.40	107.90	6.00	2288.00	239.12	3190.37	15.18
Q	88.8	7.80	1.52	2.47	16.16	1.95	0.56	16.90	44.88	92.24	4.98
R	165.6	80.60	11.04	23.27	12.88	6.76	8.00	7.80	70.24	220.59	5.31
S	216.0	5.59	3.60	0.52	45.60	5.07	5.60	10.40	139.60	215.98	9.87
T	301.6	15.60	4.56	0.65	37.60	24.70	1.60	15.60	191.28	291.59	7.38
U	468.0	30.42	91.20	2.08	18.24	9.88	46.40	88.40	182.48	469.10	8.94
V	8.8	0.26	1.44	0.13	0.16	0.26	0.88	0.00	4.88	8.01	6.84
W	341.6	22.88	17.92	1.43	48.80	18.59	22.40	17.55	204.96	354.53	7.08
X	34.4	4.68	4.00	0.52	1.92	1.56	2.80	5.20	12.72	33.40	8.58

Heavy

Min. : 4.980

1st Qu.: 7.103

Median : 8.745

Mean : 9.859

3rd Qu.:10.725

Max. :19.920

Correlations : strong → difficult context for ordinary multiple LS regression

	COND	SiO2	Na	K	Ca	Mg	Cl	SO4	HCO3	Sum	Heavy
COND	1.00										
SiO2	-0.10	1.00									
Na	0.04	0.26	1.00								
K	0.10	0.84	0.03	1.00							
Ca	0.98	-0.14	-0.11	0.08	1.00						
Mg	0.97	-0.11	-0.05	0.11	0.95	1.00					
Cl	-0.04	0.07	0.74	-0.08	-0.16	-0.11	1.00				
SO4	0.96	-0.12	-0.08	0.11	0.99	0.93	-0.14	1.00			
HCO3	0.24	0.06	0.09	-0.09	0.20	0.23	0.13	0.15	1.00		
Sum	0.96	-0.07	-0.01	0.09	0.97	0.93	-0.07	0.96	0.41	1.00	
Heavy	0.89	-0.23	0.02	-0.03	0.84	0.89	-0.01	0.78	0.31	0.82	1.00

Principal Component Analysis (PCA)

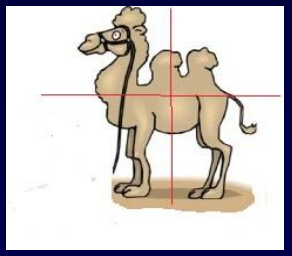
The aim : How to ‘see’ in spaces of large dimensions (>3) ?
Keywords : Exploration, latent variables, principal components.

The tool : New uncorrelated variables of max variance, in few number, called principal components or latent variables, to remove the noise and explore the data through bi-dimensional maps.

- p variables (continuous) $X^1, \ldots, X^p$
- n observations
- one data matrix : X n x p, r = rank(X)

- k uncorrelated latent variables or principal components $t^1,\ldots,t^k$

$$t = Xw = x^1 w_1 + \ldots + x^p w_p$$



1) Iterative algorithm : the principle of « largest inertia »

(1)

$$t = \max_{\|w\|^2=1} \mathrm{var}(Xw).$$

(2) once t obtained, LS regressions are made, called «partial», that are orthogonal projections of the p variables on t

$$\hat{X} \leftarrow LS(t, X) = \left\{ LS(t, x^j) \right\}_{j=1,\ldots,p}$$

(3) compute new pseudo-variables : the residuals

$$X \leftarrow X - \hat{X}$$

go to (1) to compute the next t .

When stopping, $k \leq r$, the last residual is neglected as noise.

2) A direct solution of all principal components:

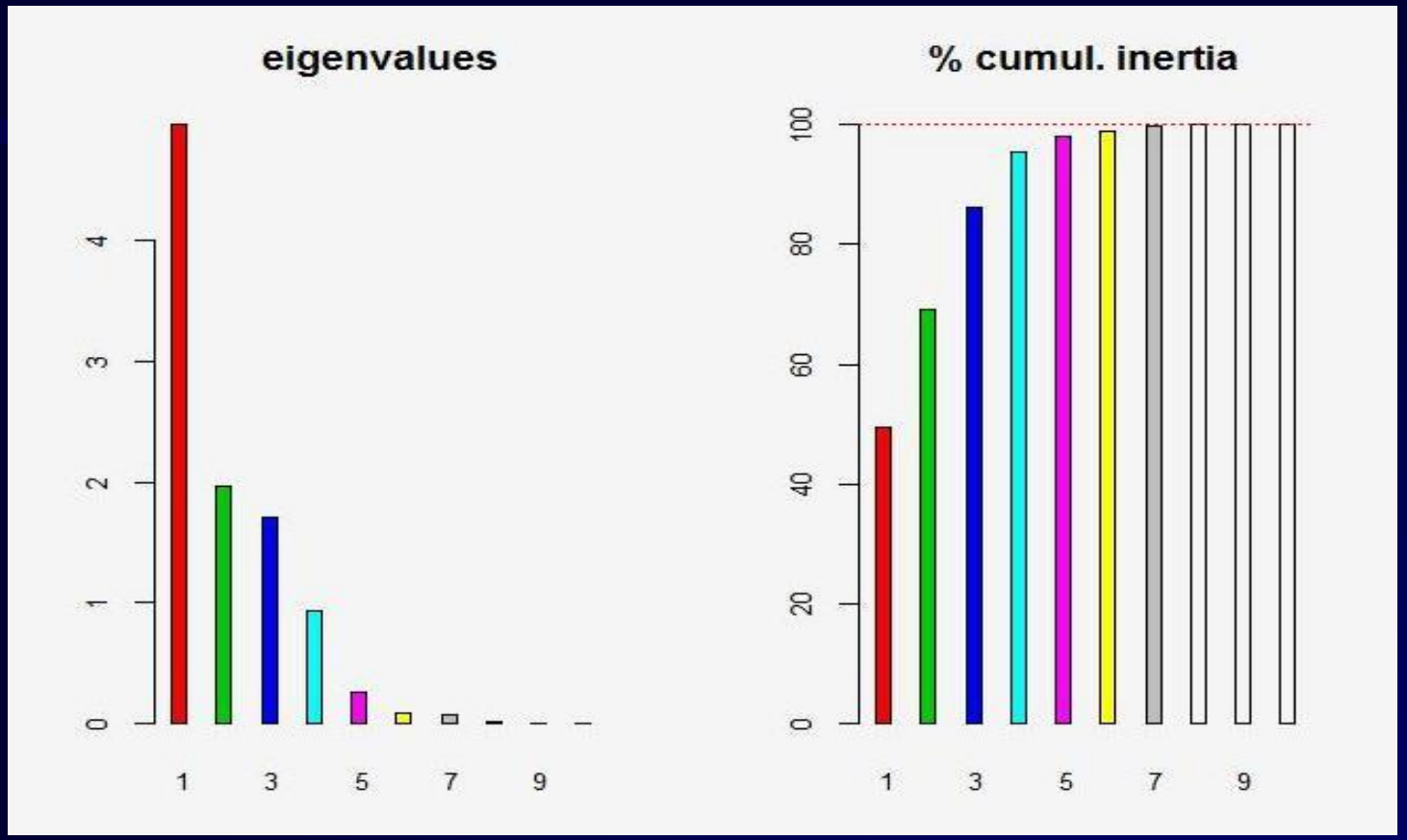
The vectors $w^1,\ldots,w^r$, solutions of 1), are the eigenvectors of the pxp matrix of correlations

$$V = \frac{1}{n} X' X$$

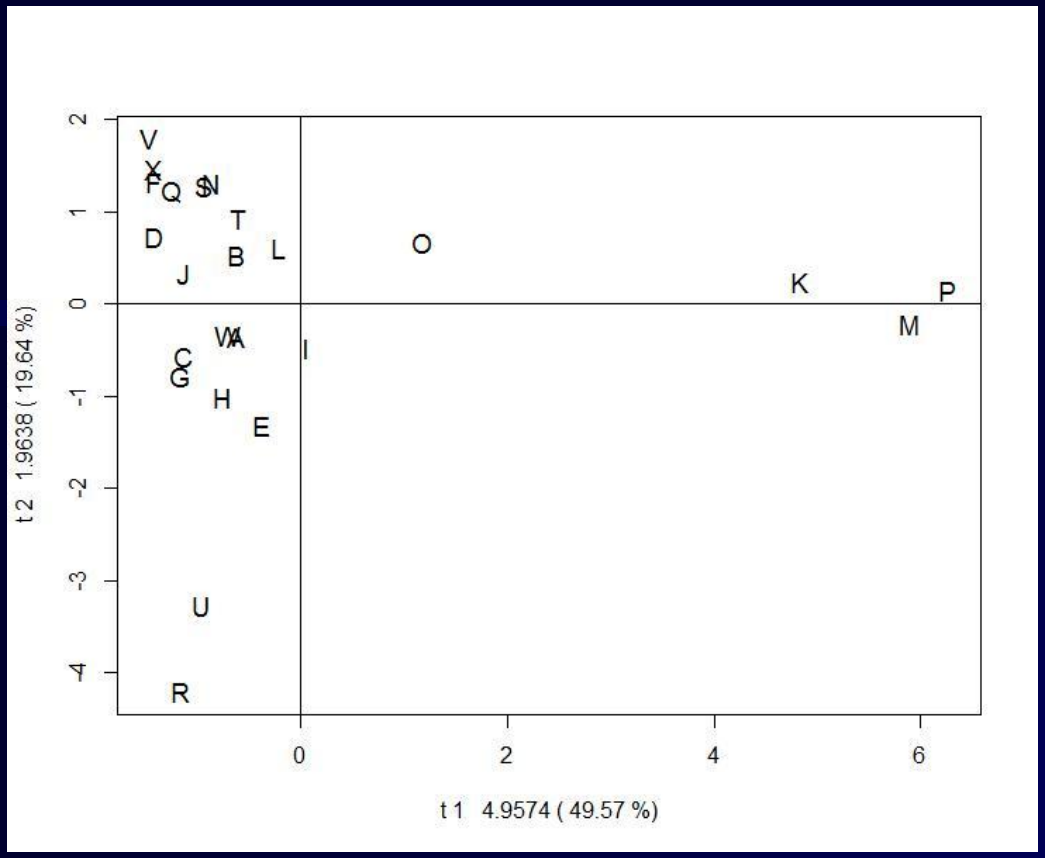
whose eigenvalues $\lambda_1,\ldots,\lambda_r$, decreasingly ordered, are the variances of the latent variables $t^1,\ldots,t^r$.

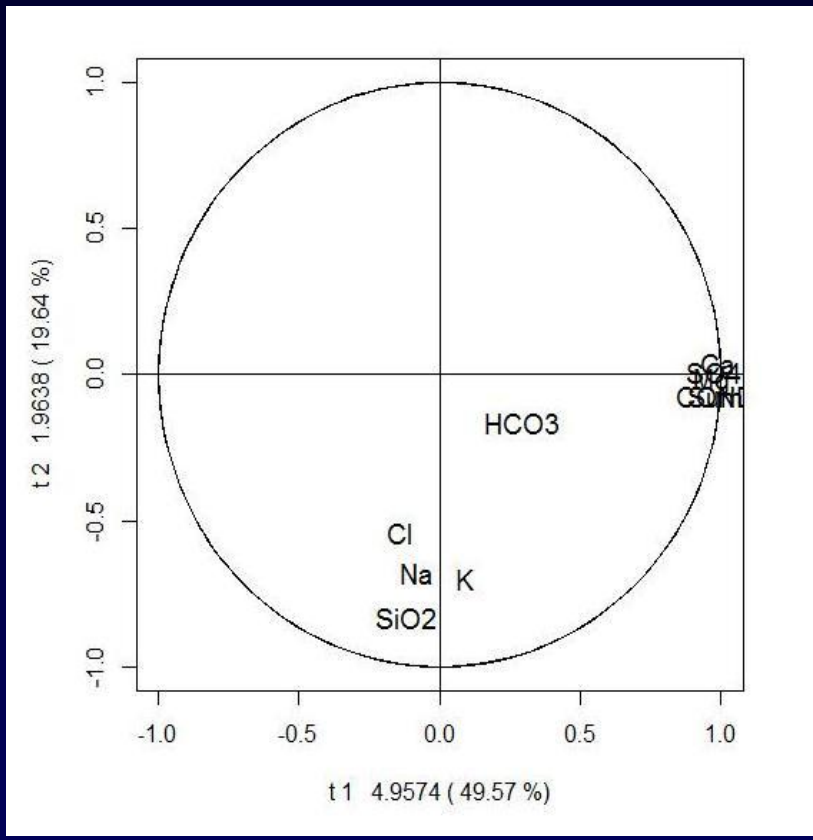
So, the inertia of the projected points: $\mathrm{var}(t^i) = \lambda_i$

Going back to the Juice data to select the best k:



Examinating the cumulated sum of the eigenvalues :
The decision is k = 3 (more than 80% of the total inertia).

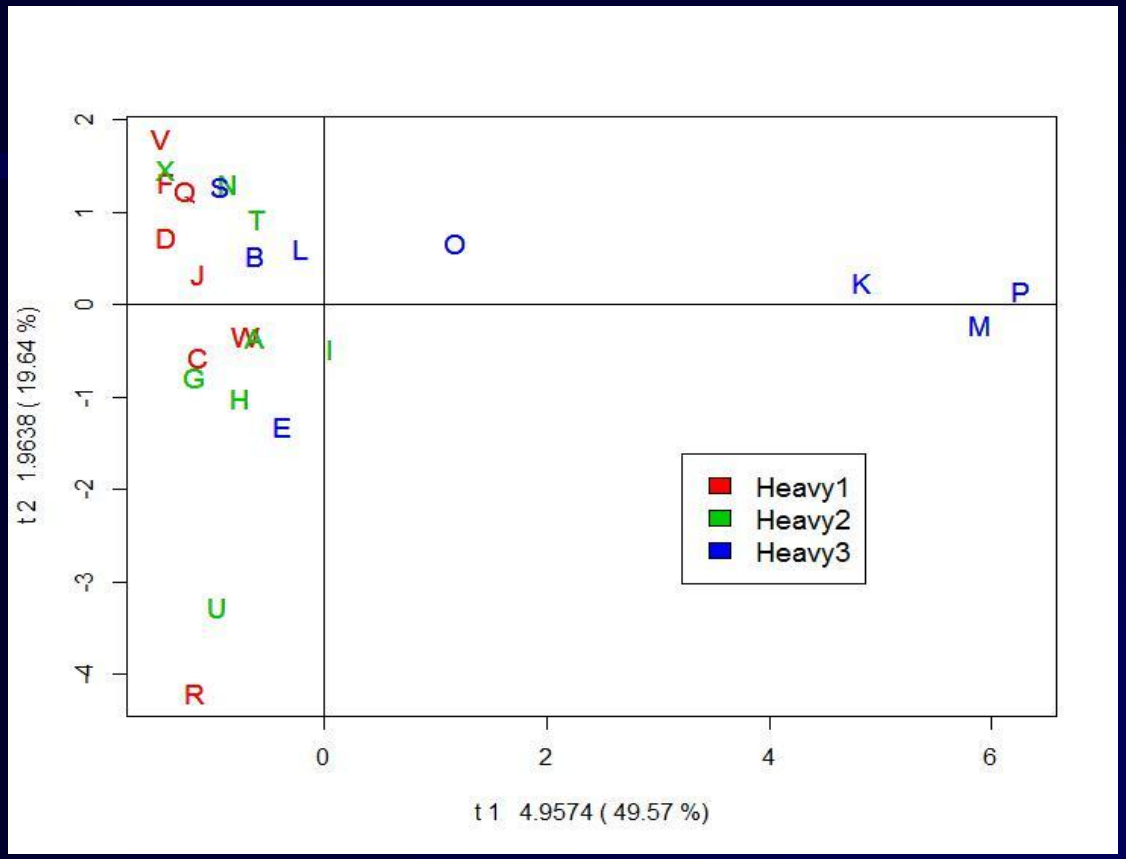


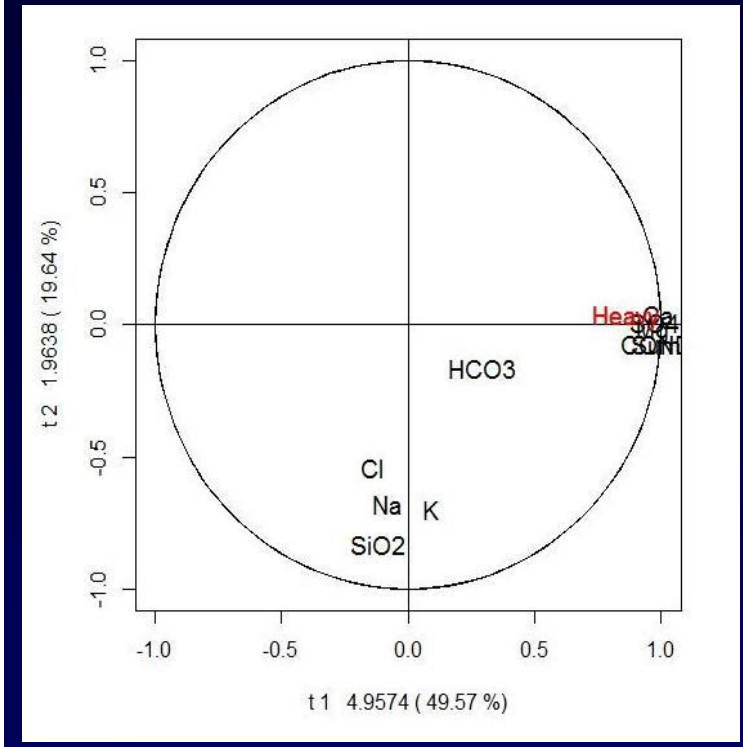


A swift way to detect the variables well represented by their projections: those that are close to the correlation circle.

HCO3, Cl, Na and K need the third axis to be well represented

How to introduce external information coming from Heavy ?





Two ways of doing :

1) Project supplementary variables (Heavy) on the components
2) Put colors on the observations according to the levels of a categorical variable, Heavy weak (1), mean (2), strong (3) .

Linear Partial Least-Squares regression (PLSL)

The aims : Explore the relationship between two data sets,
Produce a linear model of prediction (PLSL).

The tool: Latent variables from X, of max covariance with Y.

- p predictors (continuous or categorical) $x^1, \ldots, x^p$
- q responses (continuous or categorical) $y^1, \ldots, y^q$
  - continuous : regression model
  - categorical variables : classification model
- Learning data: 2 matrices : X nxp, r=rank(X), and Y nxq .

- k uncorrelated latent variables or components $t^1, \ldots, t^k$

$$t = Xw = x^1 w_1 + \ldots + x^p w_p$$

- Iterative algorithm (for simplicity, one response only)

(1) $$t = \max_{\|w\|^2 = 1} \operatorname{cov}(Xw, y).$$

(2) t obtained, « partial » LS regressions are made on t

$$\hat{X} \leftarrow LS(t, X) = \left\{ LS(t, x^j) \right\}_{j=1,\ldots,p} \qquad \hat{y} \leftarrow LS(t, y)$$

(3) compute new pseudo-variables (residuals)

$$X \leftarrow X - \hat{X} \qquad y \leftarrow y - \hat{y}$$

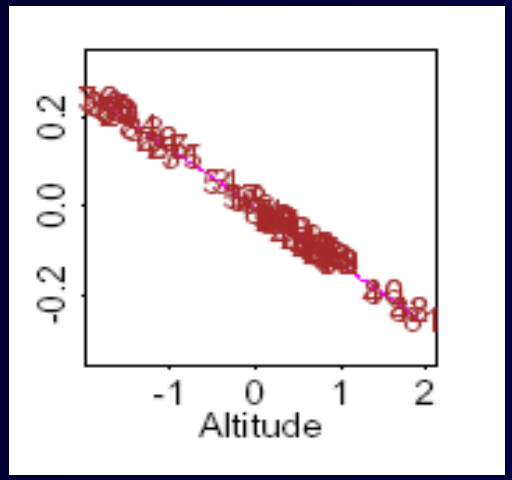
go to (1) to compute the next t .

- The final PLSL model on the retained k PLS latent variables

$$\hat{y}_{(k)} = c_1 t^1 + \ldots + c_k t^k$$

$$\hat{y}_{(k)} = \beta_k^1 x^1 + \ldots + \beta_k^p x^p$$

$\beta_k^i x^i$ is the linear function of x^i representing



the main effect of x^i on the

response y .

To summarize : PLSL (X ,Y)

Statistical properties of PLSL

- if k = r, the PLS model is identical to the multiple L-S regression of each of the q responses on X, if they exist (matrix X of full column rank r=p)

$$PLS(X,Y) \equiv \{LS(X,y_j)\}_{j=1,\ldots,q}$$

- If Y = X , PLS is identical to PCA

$$PLS(X, Y=X) \equiv PCA(X)$$

Domain of efficiency of PLSL

Large datasets .

Against the curse of the dimension, large number of predictors (p>n)

Against strongly correlated X variables.

Limits of the PLSL linear model

Nonlinearities.

Interactions between predictors

Choosing the dimension of the model : k ?

1) External validation

At disposal: One other data set on the same variables (X_s,Y_s) measured on N new observations.

a) We dispose of r PLS possible models built on (X,Y)

b) k is obtained as the dimension giving the smallest Mean-Squared Error when predicting Y_s by X_s.

2) Internal Validation by Cross-Validation (C-V)

One round of cross-validation involves partitioning the sample of data into two complementary subsets, one subset to build PLS models, the other subset to evaluate these models.

To reduce variability, multiple rounds of cross-validation are performed using different partitions

Example: The leave-one-out method :
One (Xi,yi) observation is “out” at a time to predict yi by Xi by using PLS models built on the n-1 remaining data.

The retained dimension k is that corresponding to the smallest Predictive Error Sum of Squares, the PRESS.

Tuning parameter: the proportion of observations “out” at a time.
To reduce the computing time when large datasets, 10% are accepted.

3) Internal validation by Generalized Cross-Validation (GCV)

GCV is a swift surrogate of CV where γ (≥0, defaulting to 1) controls the penalty to var(e), K is the number of predictors .

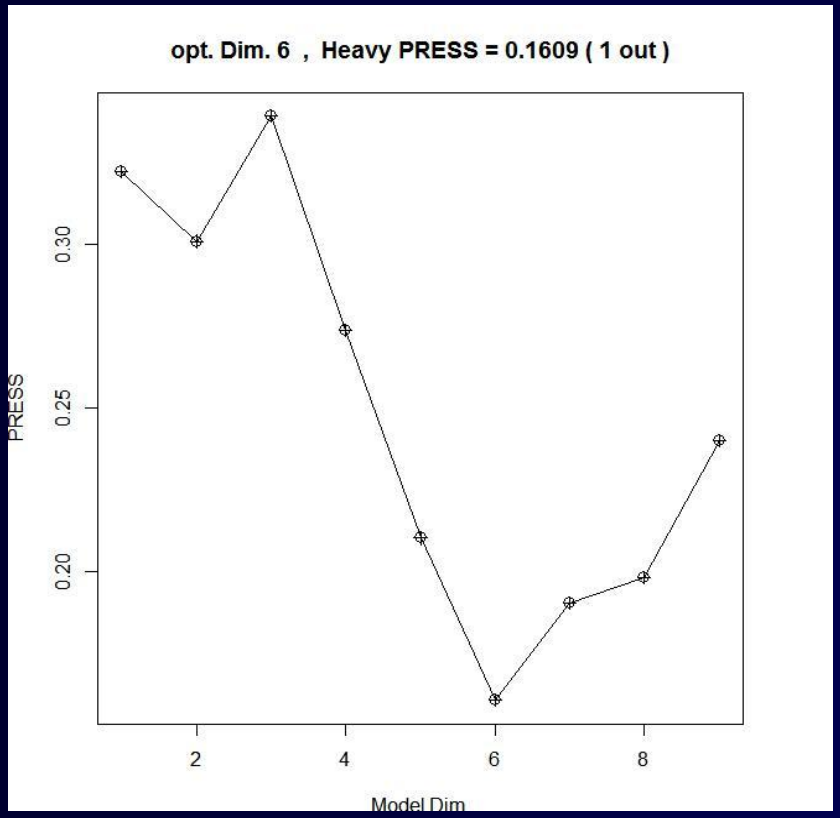
$$e = y - \hat{y}$$

$$GCV(\gamma, K) = \mathrm{var}(e)/(1 - \gamma K/n)^2$$

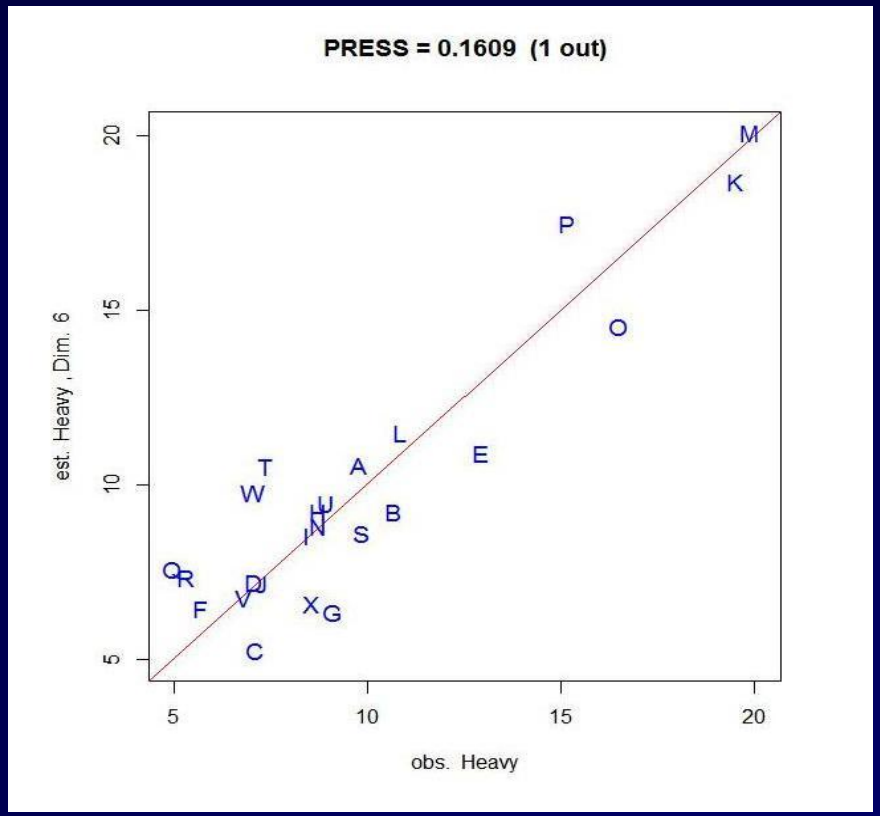
PLSL Cross-Validation on the juice dataset : 1 out at a time

optimal dimension: 6 components

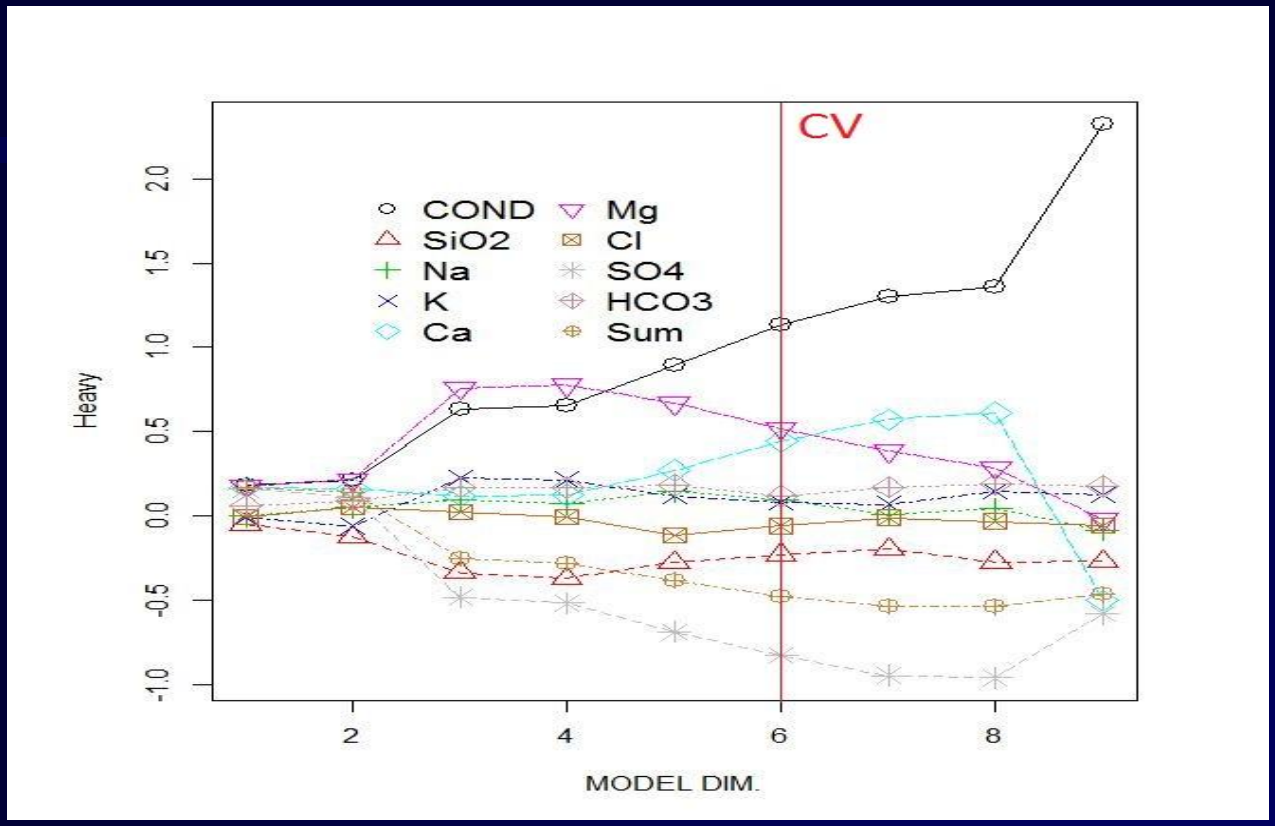
PRESS plot



(estim. Heavy , obs. Heavy)



Evolution of the PLSL β's according to model dimensions:



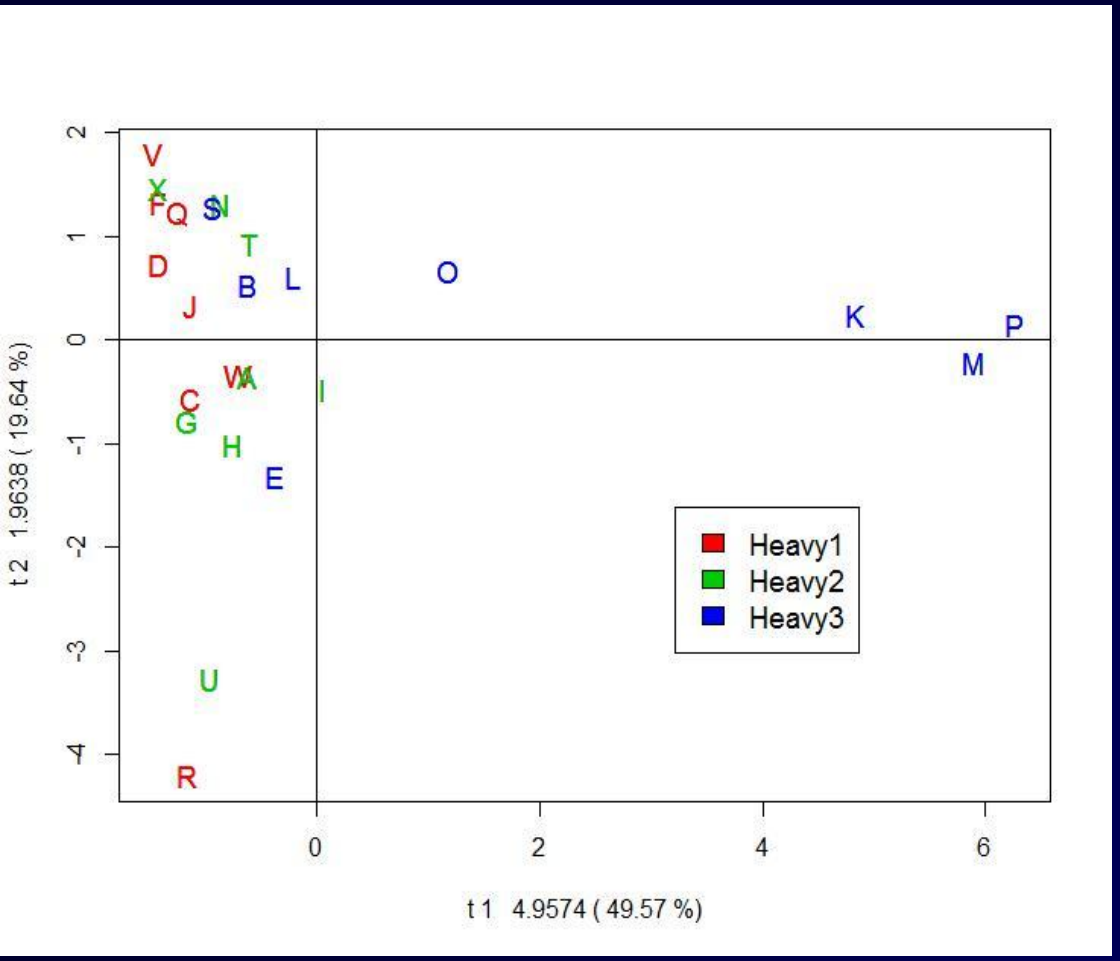
The β's at dimension 6: On standardized var. and on initial var.

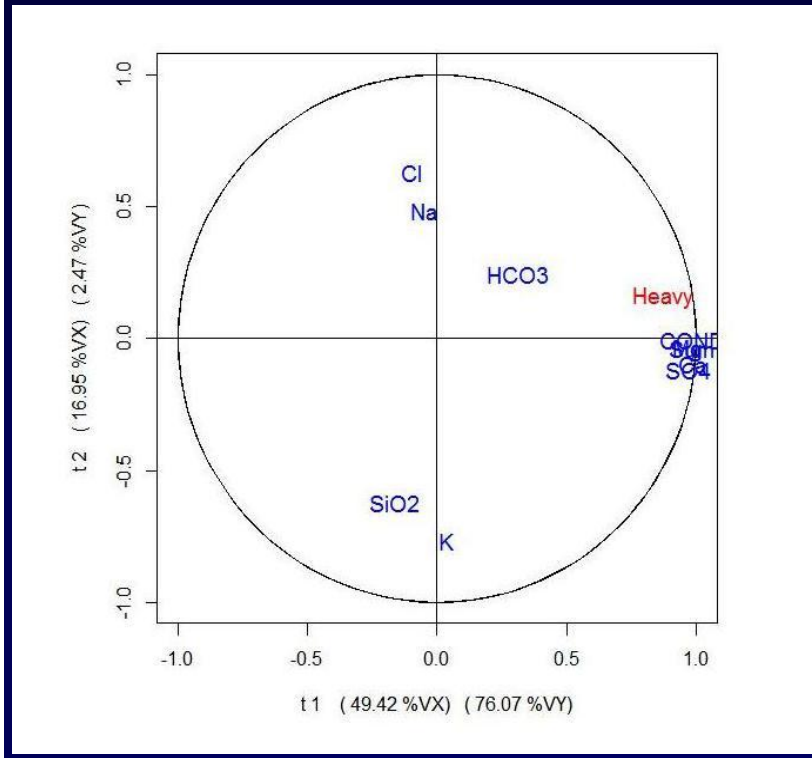
	COND	SiO2	Na	K	Ca	Mg	Cl	SO4	HCO3	Sum
Heavy	1.13	-0.23	0.09	0.08	0.45	0.52	-0.06	-0.82	0.12	-0.47

☹

	Const	COND	SiO2	Na	K	Ca	Mg	Cl	SO4	HCO3	Sum
Heavy	6.558	0.009	-0.057	0.021	0.076	0.013	0.054	-0.016	-0.005	0.002	-0.002

A look at data through (t1,t2) components plots





PLSL as a linear classifying supervisor

PLS as a classifying supervisor:

- Y is the boolean indicator matrix of the groups (q>1)

  t latent variables → discriminant variables.

- Using the learning set (X,Y) PLSL diagnoses in which class of Y to assign new X-observations.

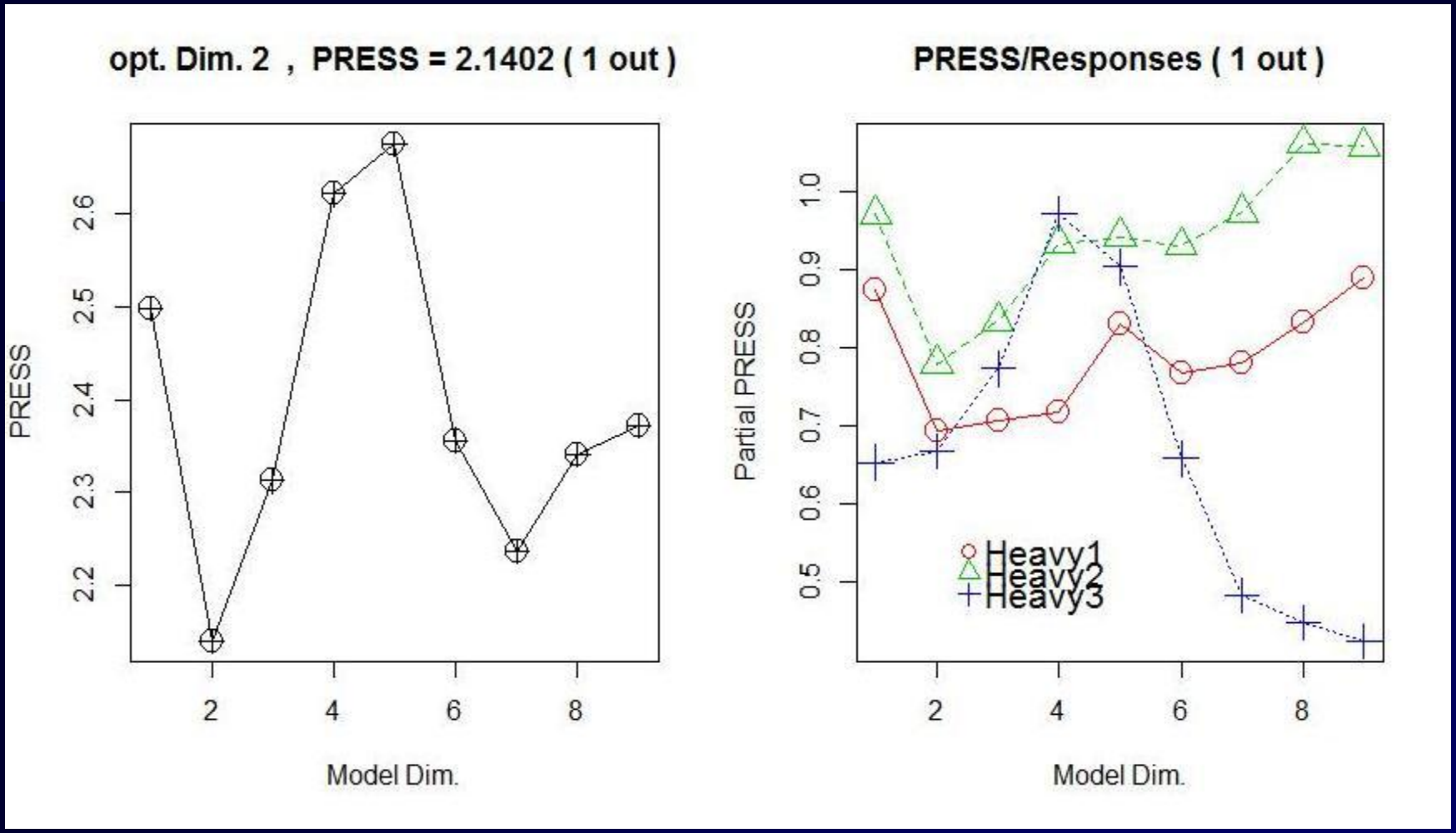
Juice dataset: Heavy is split up in 3 groups (breaks at quantiles)

g categorical (3 groups) → Y (24 x 3)

$g_i = 1$ (weak Heavy) → $Y_i = [1 , 0 , 0]$

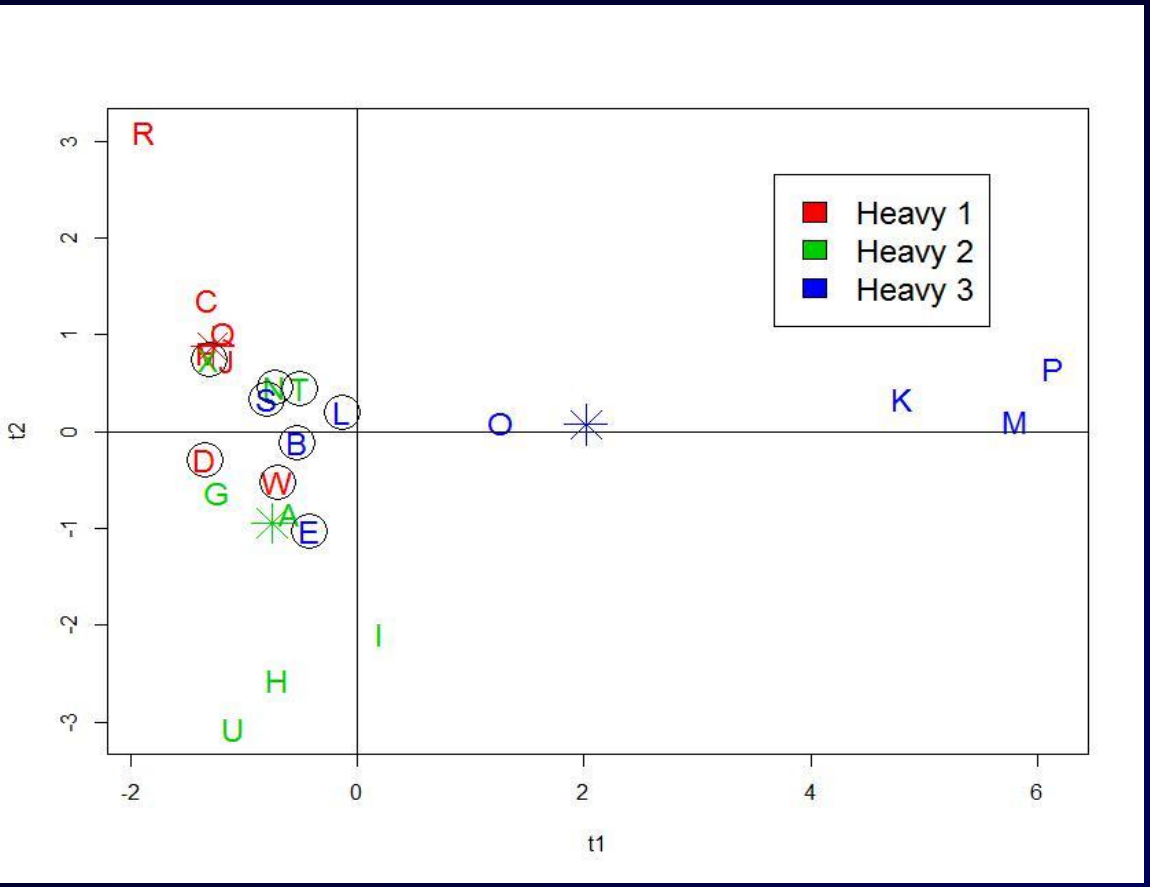
$g_i = 2$ (mean Heavy) → $Y_i = [0 , 1 , 0]$

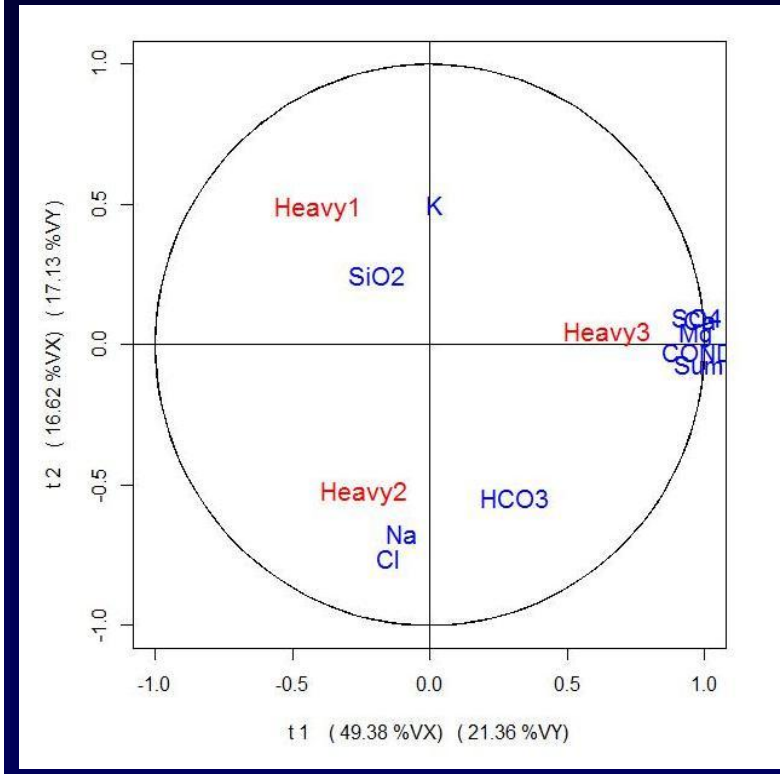
$g_i = 3$ (strong Heavy) → $Y_i = [0 , 0 , 1]$



prop=0.05 (proportion of observations → 1 ‘out’ at a time)
PRESS(0.05, 2) = 2.14 < 3 → PLS discrimination accepted

Classification using PLSL components :





Geometric rule in the $\{t_i\}_{i=1,\ldots,k}$ space for classifying observations :

a) Compute the k discriminant variables for one new observation.

b) Classify that item in the group whose centroïd (∗) is of smallest Mahalanobis distance

One way (rather optimistic) to assess the goodness of the method is to apply the classifying rule to the training set itself.

Heavy	weak	mean	strong
predicted weak	6	3	2
predicted mean	2	5	2
predicted strong	0	0	4

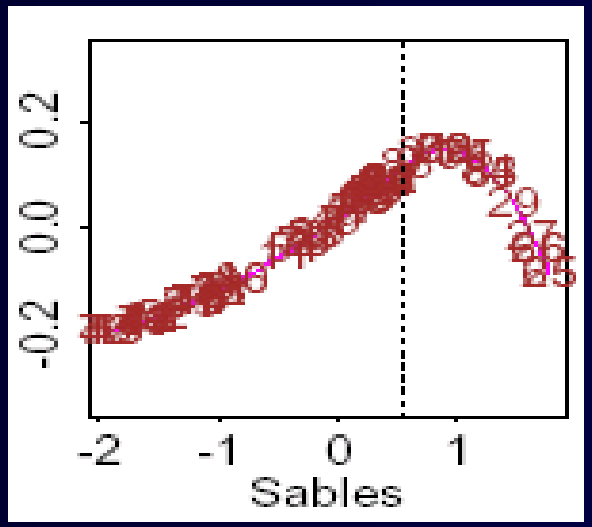
Percentage of misclassified items: 36%

Partial Least-Squares Splines (PLSS)

- Additive models through k latent variables $t^1, \ldots, t^k$

$$\hat{y}_{(k)} = s_k^1(x^1) + \ldots + s_k^p(x^p)$$

$s_k^i(x^i)$ is a nonlinear function of x^i representing the main effect of x^i on the response y : a spline



degree 2, one knot at 0.5

A spline is made of piecewise polynomials of the same degree, that join end to end at points called the ‘knots’

A spline (local piecewise polynomial function)

$$s(x) = \sum_{i=1}^{s} \beta_i B_i(x)$$

used to transformed one variable x, is a linear combination of s basis functions (the B-splines)

☞ Prediction inside the training data

s = d+1+K the dimension of the spline

d = degree of the polynomials

K = the number of knots (the junction points)

The new predictor matrix $B = [B_1 B_2 \ldots B_p]$

$n \times (s_1 + s_2 + \ldots + s_p)$ super-coding matrix

To summarize : PLSS(X ,Y)= PLSL(B ,Y)

Principal Components : depend nonlinearly on the predictors. They are splines of the same type (degree, knots) than those used to model Y.

Tuning parameters of PLSS:

- The spline for each predictor x_i
  - the degree d_i
  - the « knots» : their number K_i and their locations

- The model dimension : k (CV or GCV)

Advantages of PLSS

⇒ against colinearity of predictors

⇒ against small ratio #observations / #predictors

⇒ easy to interpret the main effects spline functions

1) PLSS regression on the juice dataset

The retained splines :

same degree 2 for all predictors

knots :

COND	SiO2	Na	K	Ca	Mg	SO4	HCO3	Sum
400	10	10	2.5	160	40	400	100	600
1600	20	40	5	400	110	1700	300	2600
	40						500	

How to choose the degree and the knots for one predictor?

Adding a knot increases the local flexibility of the spline and then the freedom of fitting the data in this area.

A campaign of tries of degrees and knots.

A balance between

Parsimony

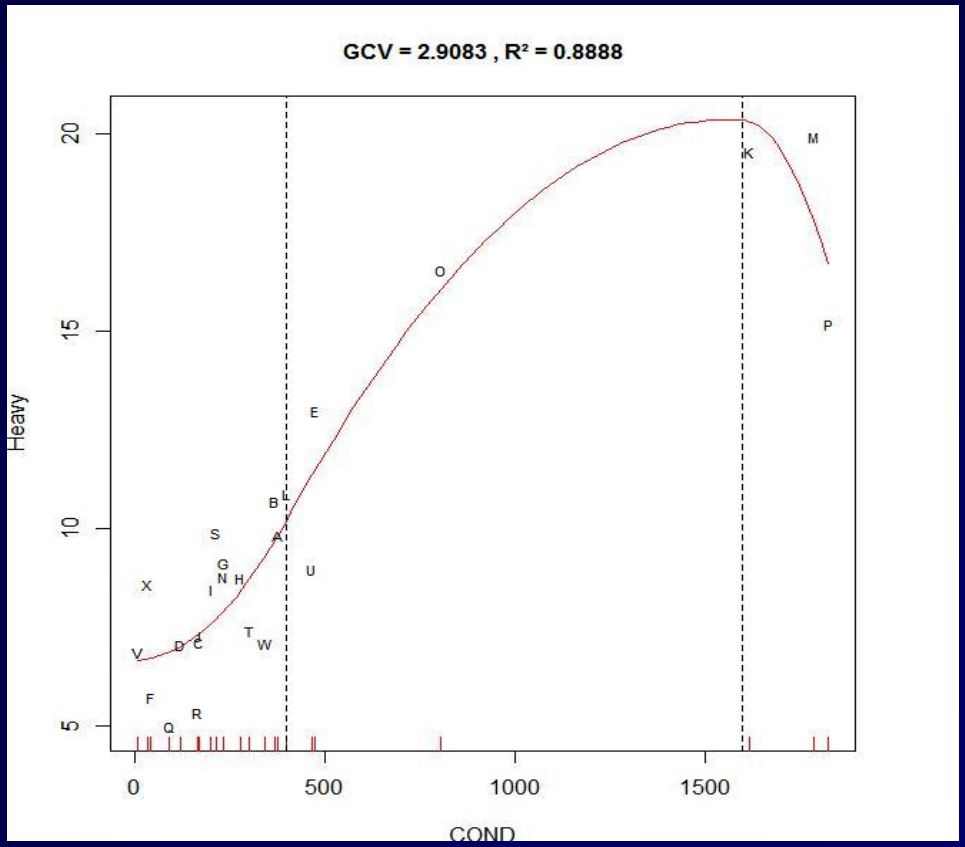
Spline dimension (s lower): degree, nb of knots

Goodness

Criterion:

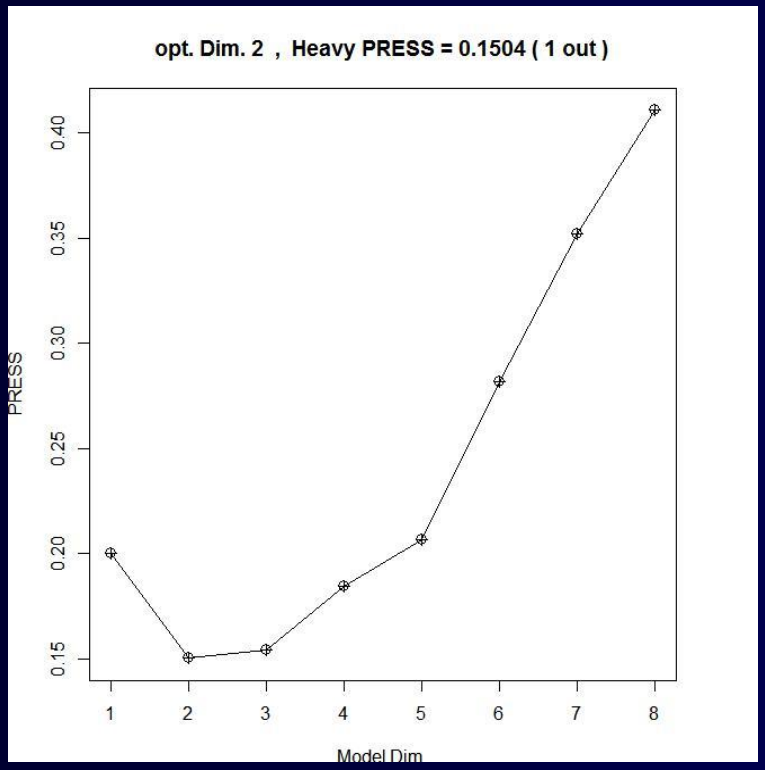
R^2 (higher) and GCV (lower)

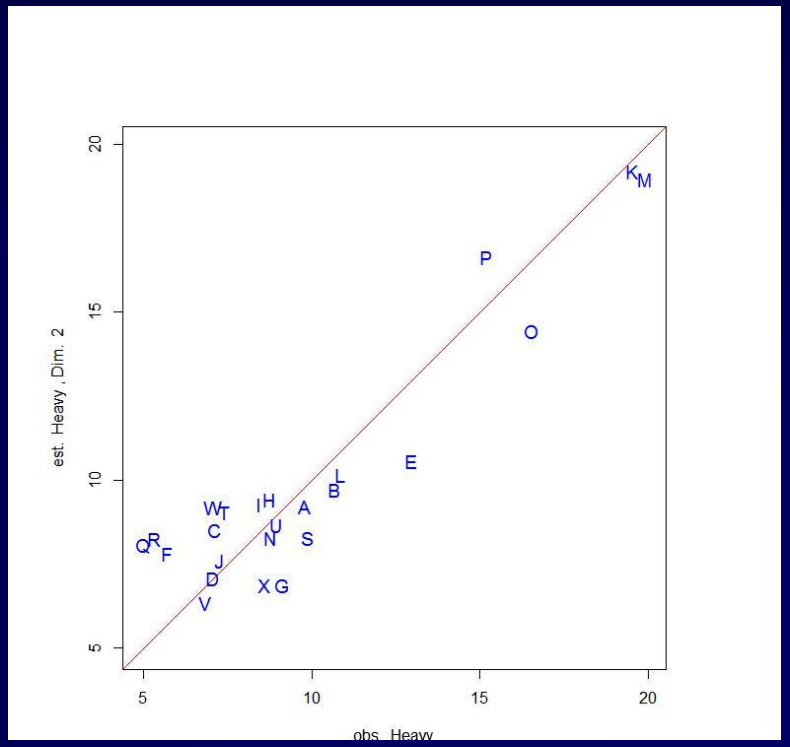
bivariate regression : L-S Spline (COND, Heavy)



Leave-one-out PRESS optimal dimension k = 2

PRESS(0.05,2) = 0.1504



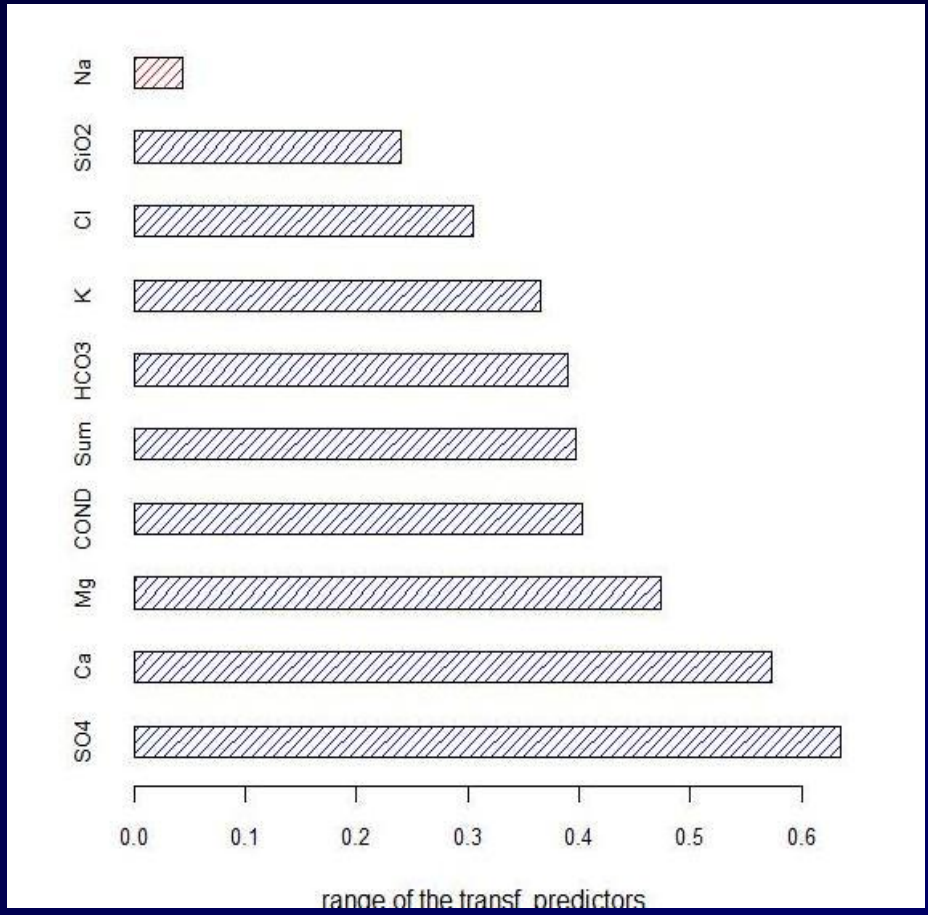


The PLSS model of Heavy with 2 components

How to order the predictors according to their influence on the response?

Criterion :

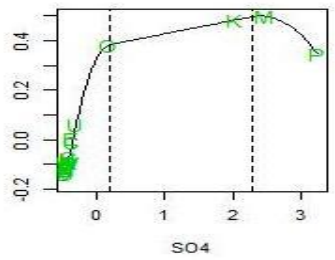
For each x^i , the range of the n values of $s_k^i(x^i)$, that are the observations transformed by the spline function.

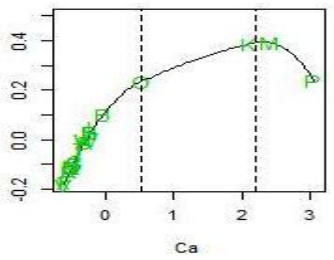


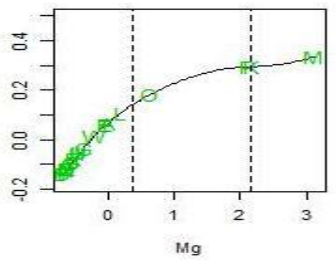
Main effects on Heavy with k = 2 components

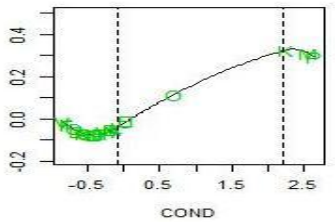
First + → -

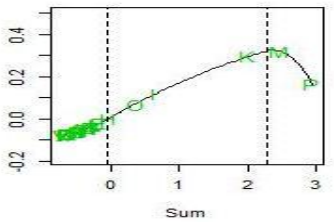
Second + ↓ -

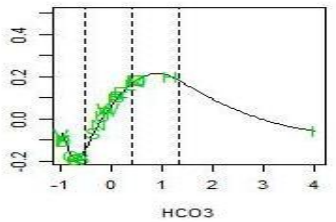


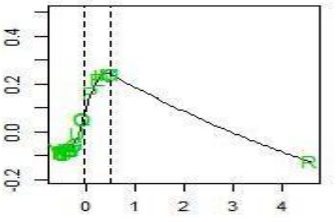


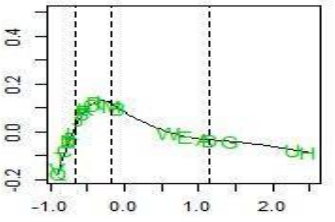


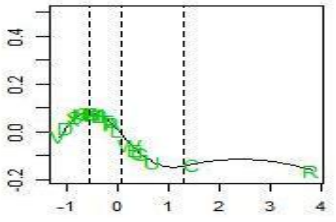


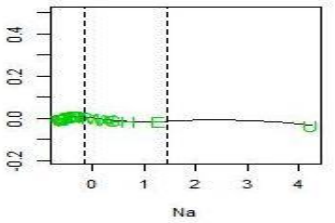




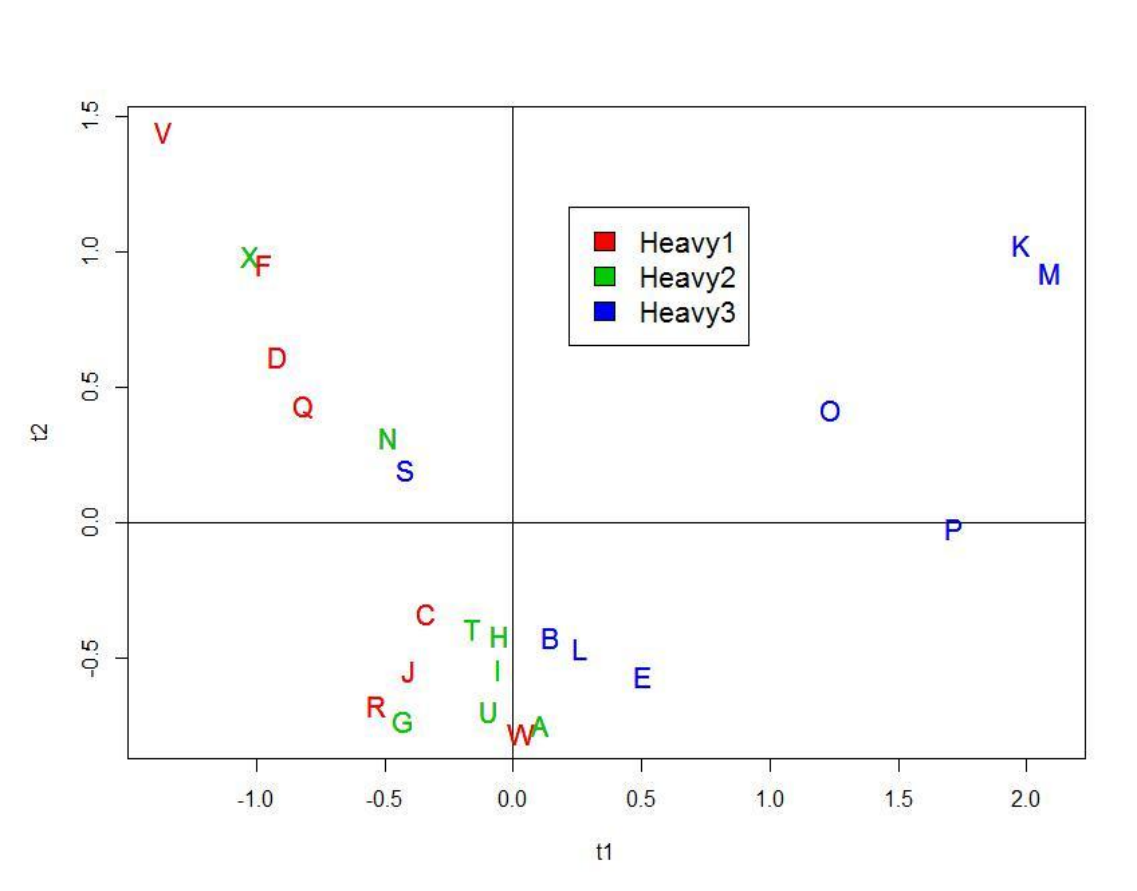


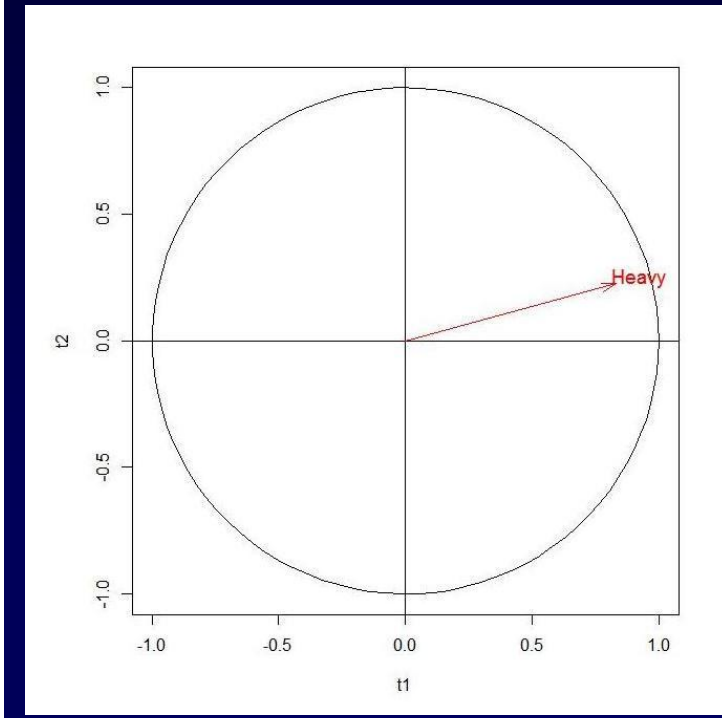






A look at data on the (t1 , t2) scatterplot :

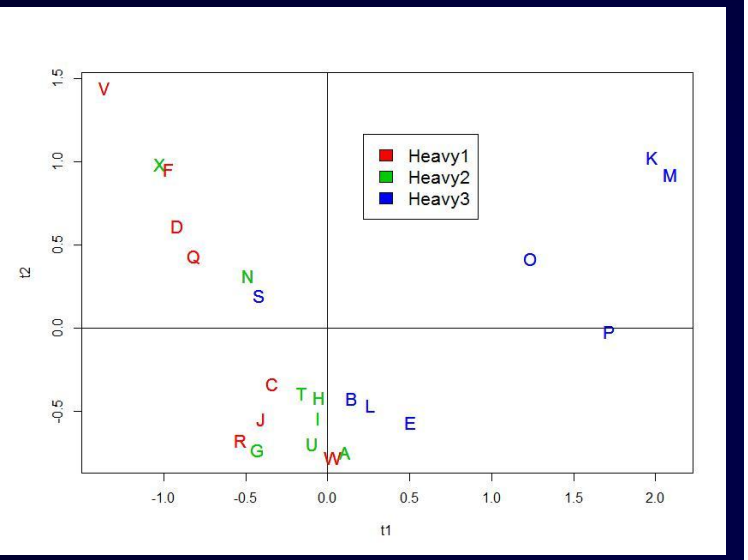


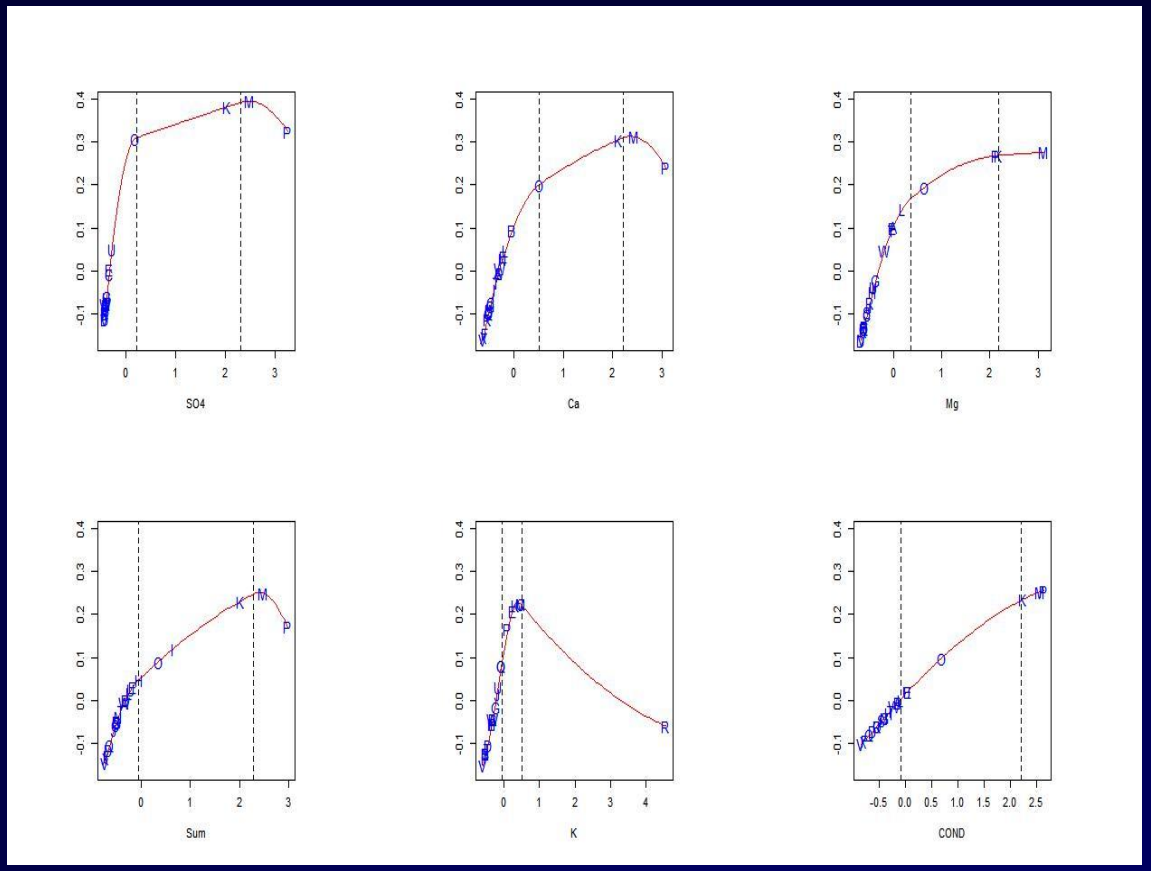


Predictors do not participate to the correlation circle

since components t_i depend nonlinearly on the predictors !

For simplicity, let us explain t_1 by the 6 first predictors only :





$t_1 > 0$

M, K, P, O from higher values of SO4, Ca, Mg, SUM and COND

B, L, E, from higher values of Ca, Mg, and mean values of K

2) PLSS discrimination of the juice data

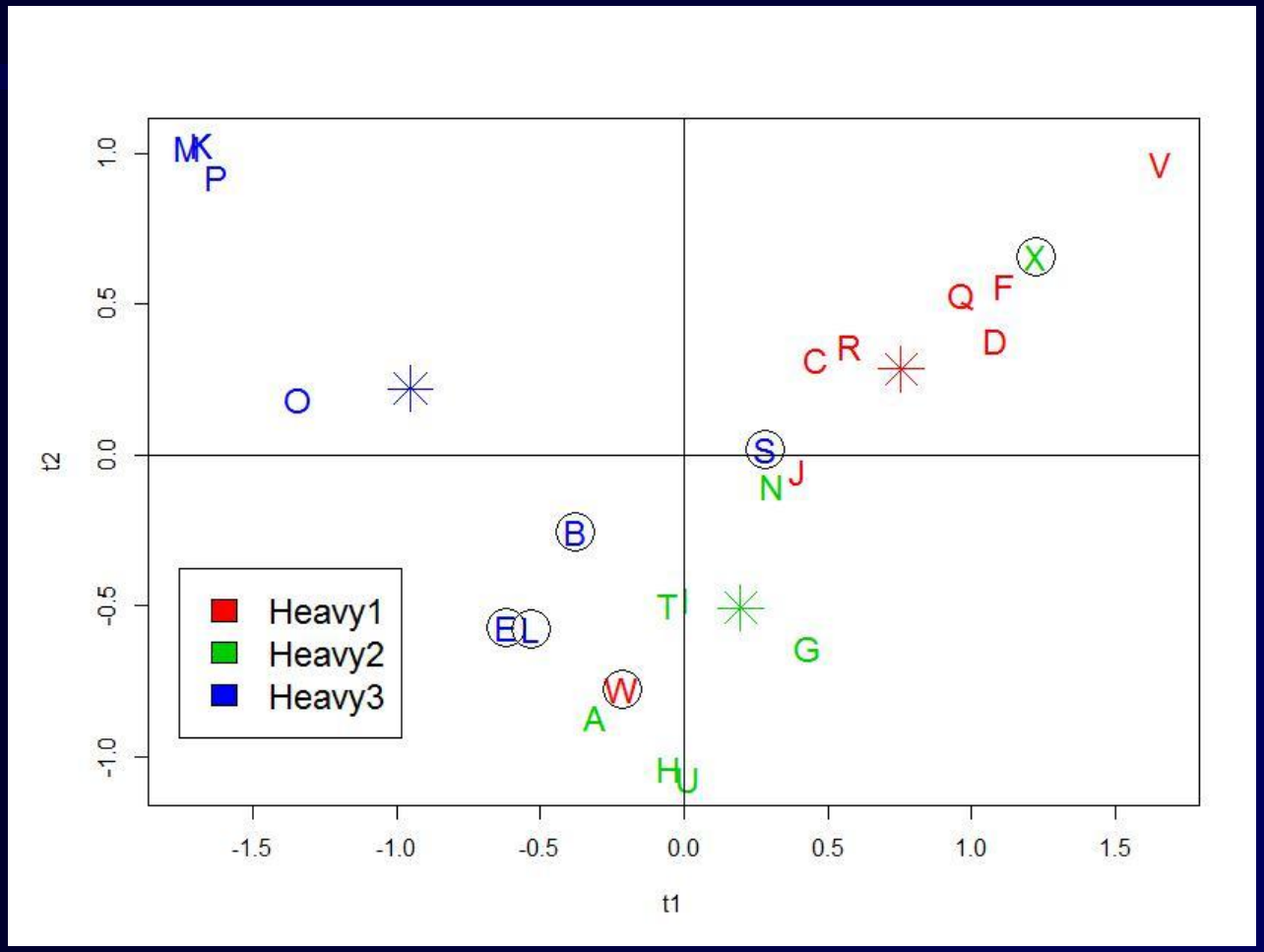
q=3 , Y= boolean indicator matrix of groups for Heavy.

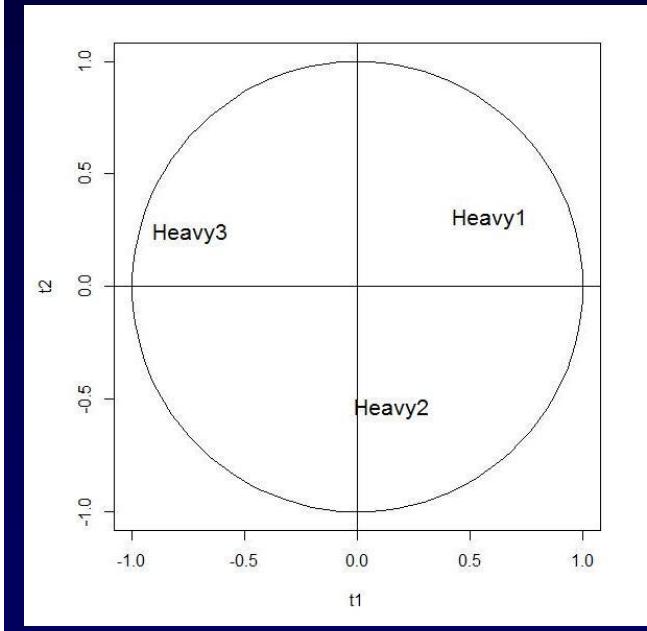
Splines : degree and knots (those used in the PLSS regression on Heavy)

Heavy	weak	mean	strong
predicted weak	7	1	1
predicted mean	1	7	3
predicted strong	0	0	4

Percentage of misclassified items: 25%

t1,t2 components scatterplot :





PLSS compared to PLSL on the Juice dataset:

Better prediction (PRESS, 0.15 against 0.16)

Better classification (misclassified 25% against 36%)

More parsimonious (2 components against 6).

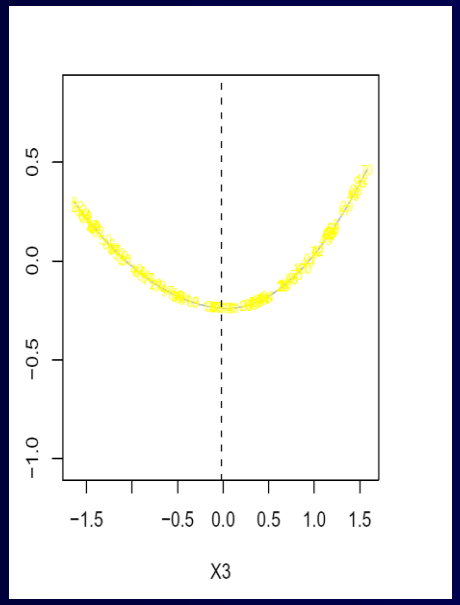
Better description of data (component plots).

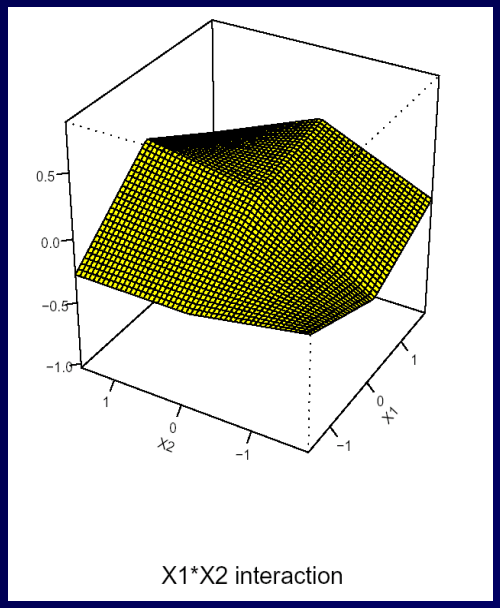
Multivariate Additive PLS Splines : MAPLSS (bivariate interactions)

- Model casted in the ANOVA type decomposition :

$$\hat{y}_{(k)} = \sum_{i=1}^{p} s_k^i(x^i) + \sum_{(i,i')\in I} s_k^{(i,i')}(x^i, x^{i'})$$

ANOVA spline functions





- The curse of dimensionality

The price of nonlinearity : expansion of the dimension of B

MAPLSS(X,Y) = PLSL(B,Y)

$B = [B_1 B_2 \ldots B_p \mid \ldots B_{i,j} \ldots]$ super- coding of X

Example : p predictors $\Rightarrow$ (p -1)p / 2 possible interactions

spline dimension = 10 for each predictor

	#columns of the design matrix B
Linear PLS	p
PLSS (main effects)	10p
MAPLSS (all bivar. interactions)	$10p + 10^2p(p - 1)/2 \sim (10p)^2/2$

Necessity of eliminating (xi,xj) non influent interactions

- 1) Automatic selection of candidate interactions :

Denote

$$CV_m(k) = PRESS(\gamma, k)$$ or $$CV_m(k) = GCV(\alpha, k)$$

each interaction i is separately added to the main effects model m and evaluated

$$CRIT(k) = \frac{R^2{}_{m+i}(k) - R^2{}_m(k)}{R^2{}_m(k)} + \frac{CV_m(k) - CV_{m+i}(k)}{CV_m(k)}$$

Rule: Order decreasingly interactions, refuse one if CRIT(k)<0

- 2) Add step-by-step ordered candidates to the main effects model, and accept a model if it significantly improves CV

- 3) Pruning step : Selection of main effects and interactions according to the range of the ANOVA functions (CV/GCV)

- Advantages of MAPLSS :
  - inherits the advantages of PLSL and PLSS
  - captures most influential bivariate interations
  - easy interpretable ANOVA function plots

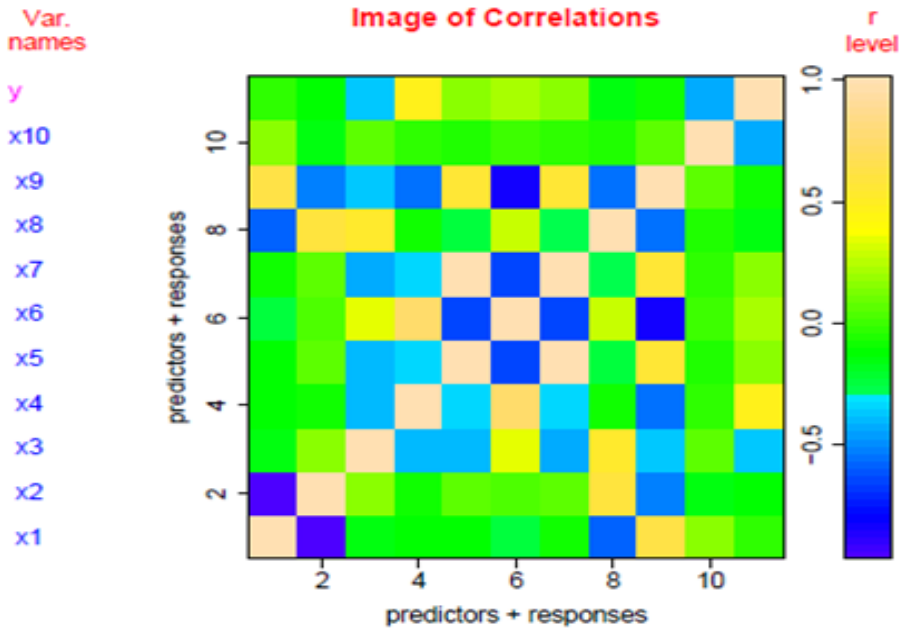
- Disadvantages of MAPLSS :
  - no higher interactions
  - no automatic selection of spline parameters

Multi-collinearity and interactions : the 'chem' data

61 observations from a proprietary polymerization process.

10 explanatory variables, $x_1, \ldots, x_{10}$ (inputs) and 1 response, y (output)

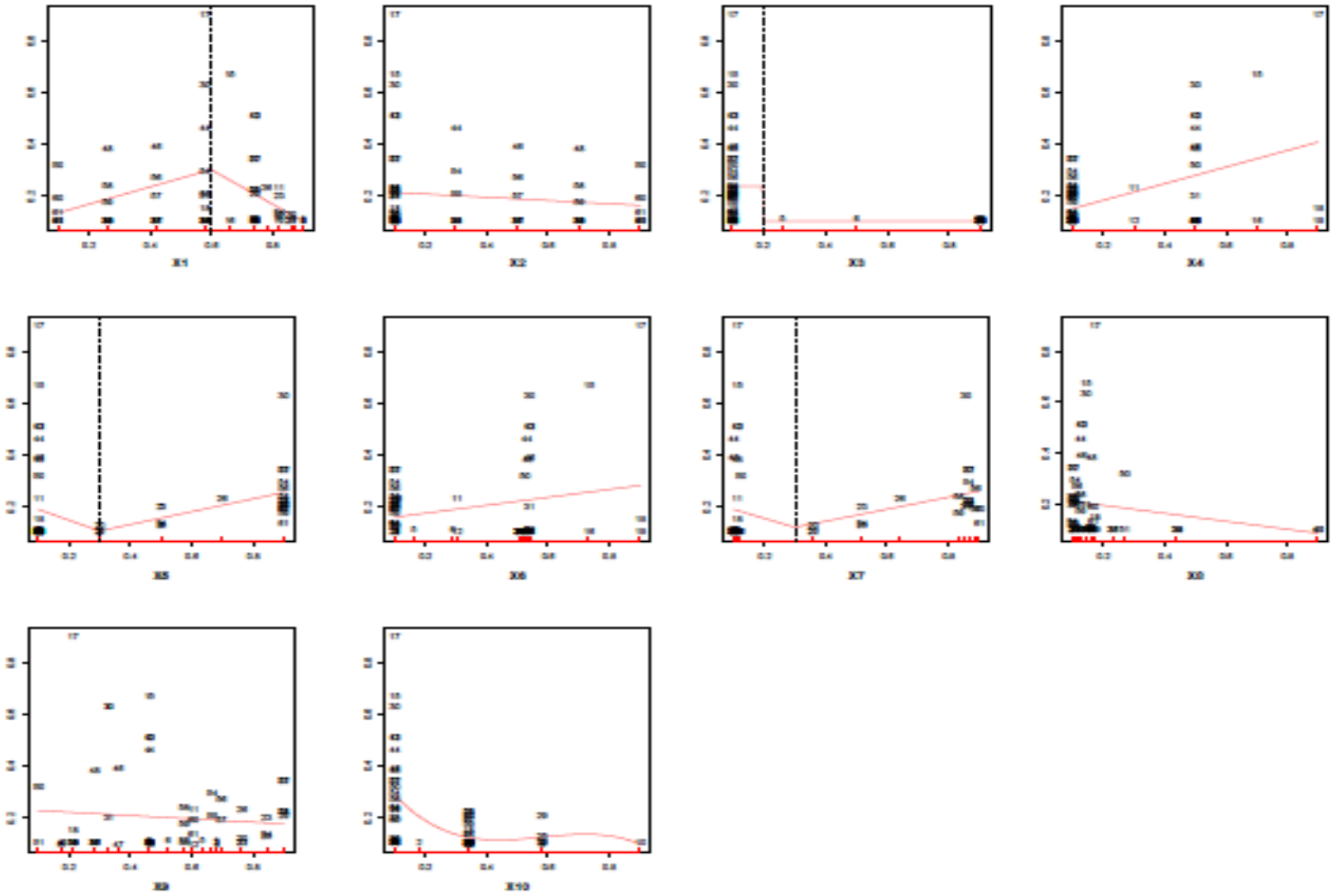
Due to the proprietary nature of the process, no more information could be disclosed.



$cor(x_5, x_7) = 1$, $cor(x_1, x_2) = -0.95$, $cor(x_6, x_9) = -0.82$

The choice for splines :

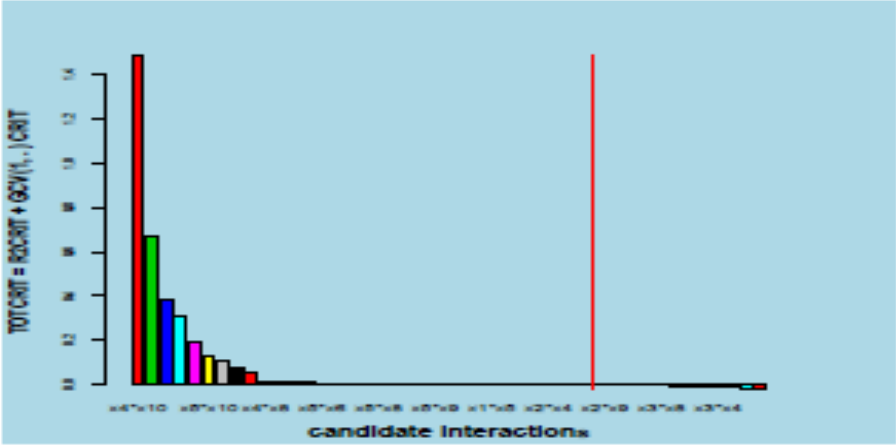
$\{(x_i,y)\}_i$ *plots of chem data smoothed with L-S splines.*



Towards a final model...

Phase 0: build a pure main effects model (PLSS)

Phase 1: select interactions candidate in decreasing order



Phase 2: Add interactions to the main effects model

candidate 1 : $x_4 * x_{10}$ ACCEPTED at 20 % relative GCV gain

	i	j	M	GCV	% rel. GCV gain
$x_4 * x_{10}$	4	10	9	0.04820853	89.22

candidate 2 : $x_6 * x_{10}$ NOT ACCEPTED at 20 % relative GCV gain

	i	j	M	GCV	% rel. GCV gain
$x_6 * x_{10}$	6	10	9	0.04757621	1.31

Continue anyway exploring (y/n)?1: n

Phase 3: Evaluate the PRESS(0.02,7)=0.0584 and ANOVA terms

Range of the transformed predictors in descending order

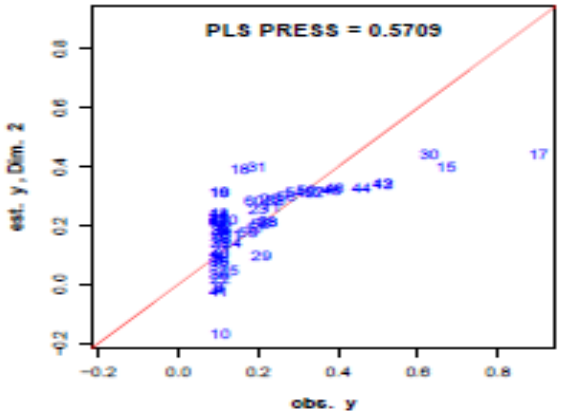
$x_4 * x_{10}$	x_{10}	x_8	x_5	x_7	x_2	x_6	x_4	x_9	x_1	x_3
3.8238	1.7266	0.6202	0.5647	0.4616	0.4163	0.266	0.2106	0.188	0.1457	0.0762

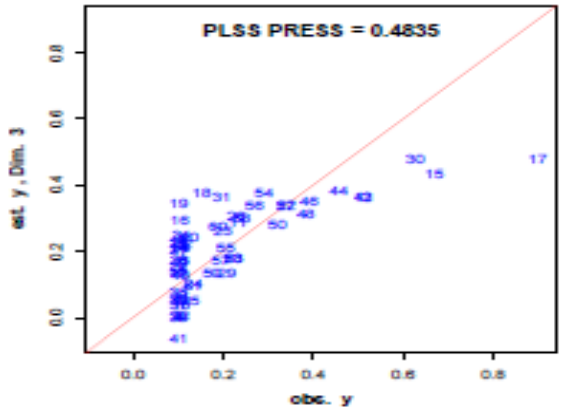
Phase 4: Prune low ANOVA terms. Final PRESS(0.02,6)=0.0569

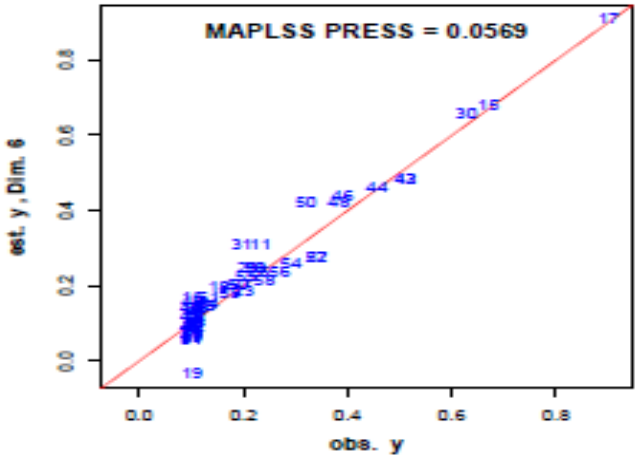
$x_4 * x_{10}$	x_8	x_2	x_5	x_7
5.11868	0.55858	0.55846	0.53523	0.42331

Comparison between PLSL, PLSS and MAPLSS

Leave-one-out predicted versus observed y samples.

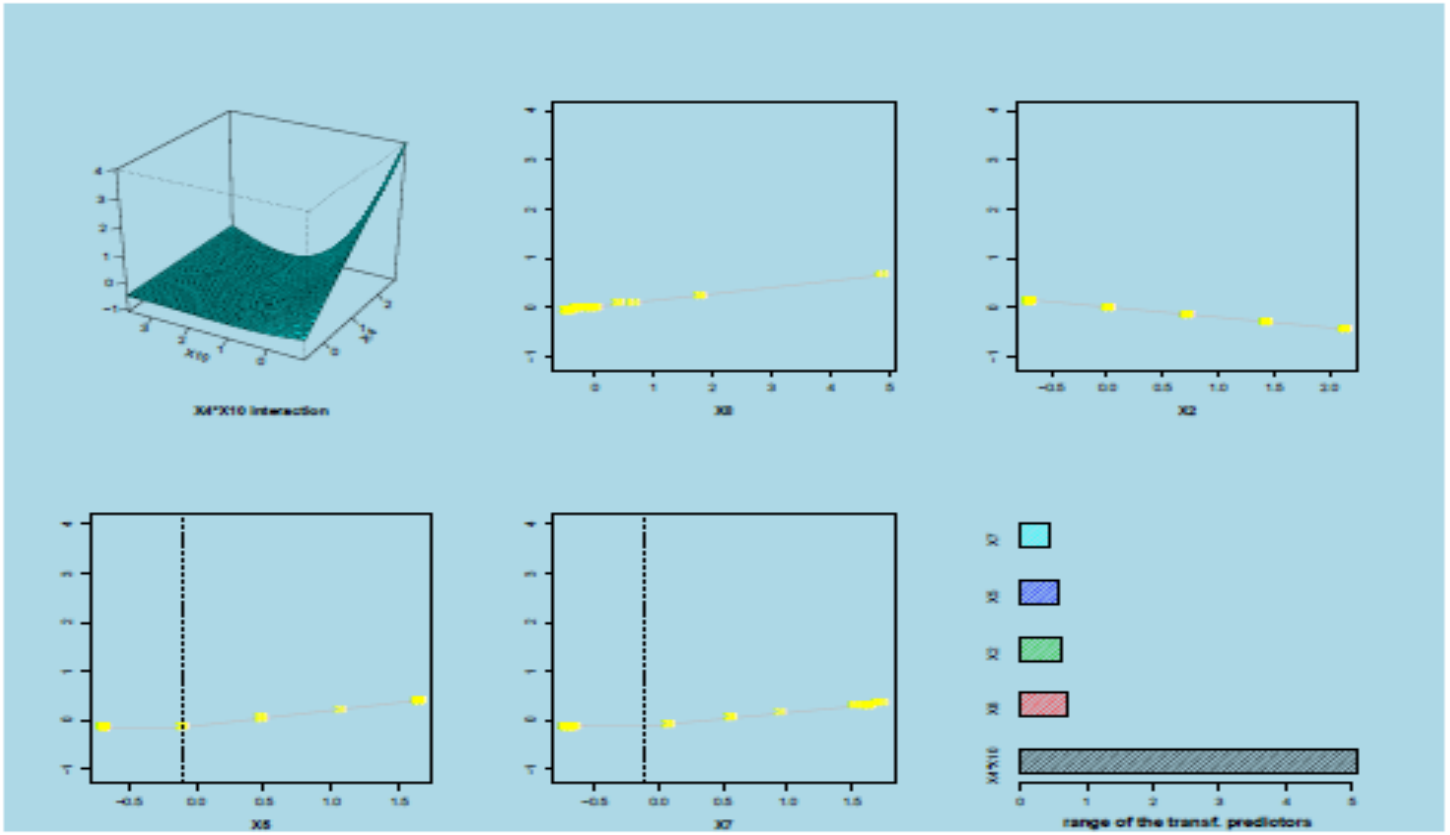






The MAPLSS model

y-ANOVA plots ordered from left to right and up to down according to the vertical ranges.



Seminal Papers

- H. Wold, Soft modeling by latent variables, the nonlinear iterative partial least-squares apprcoah, in J. Gani (Ed.) , Perspectives in Probability and Statistics, (Papers in Honour of M.S. Bartlet). Academic Press, London, 1975.

- J. F. Durand. Local Polynomial Additive Regression through PLS and Splines: PLSS, Chemometrics and Intelligent Laboratory Systems 58, 235-246, 2001.

- J. F. Durand and R. Lombardo. Interactions terms in nonlinear PLS via additive spline transformations. « Between Data Science and Applied Data Analysis », Studies in Classification, Data Analysis, and Knowledge Organization . Eds M.Schader, W. Gaul and M. Vichi, Springer, 22-29, 2003