Multivariate Data Analysis:

A nonlinear approach through

PLS and Regression Splines

J.F. Durand

Laboratoire de Probabilités et Statistique,
University of Montpellier II, France

E-Mail: jfd@helios.ensam.inra.fr

Summary

III. A Nonlinear Fuzzy Coding Approach through Regression Splines

1. What are regression splines?

- Two sets of univariate basis functions
- The attractive B-splines family for coding
- Bivariate regression splines
- The Least-Squares Splines (LSS)

2. PLS through Splines (PLSS): exploring and modelling data nonlinearly

- Multi-collinearity, nonlinearities and outliers: the orange juice data
- The main effects additive PLSS model
- A nonlinear look at data
- PLSS with bivariate interactions

IV. Open problems with PLS and Splines

V. Bibliography

Section I.

Introduction

The aim of this course is twofold and presents

- The general approach of dimension reduction in linear Data Analysis through PLS
  - for exploring proximities in the row-column spaces
  - for prediction purpose
- A nonlinear extension of PLS aiming at revisiting D.A.

Main idea:

Capture non-linearities through the linear framework of splines

The price to be payed:

Expansion of the column dimension of the design matrix

Key words and inspiring papers:

Data Analysis methods and Euclidian spaces, [5, ECAS, 1987]

Coding approach in Data Analysis, [7, Gifi,1990]

Partial Least-Squares regression, PLS, [13, 14, H. and S. Wold, 1966,1983]

Multivariate Adaptive Regression Splines, MARS, [6, Friedman, 1991]

PLS through Splines, PLSS, [2, Durand, 2002]

Section II.

Dimension Reduction

in

Linear Multivariate Data Analysis

1. Euclidian spaces for column-variables and for rows

Denote

$$X = [X_i^j] = [X^1 ... X^p] = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \quad \text{an } n \times p \text{ matrix of data}$$

from the column-variable's side

- $(\mathbb{R}^n, D)$ the Euclidian space for sampled variables
- $D = diag(p_1, \ldots, p_n)$ the matrix of the statistical weights
- $\mathbf{1} = (1, ..., 1)'$ the column vector of n repeated 1

We suppose that the variables x, y ... are D-centered, then

- $\overline{x} = \mathbf{1}'Dx = \sum_i p_i x_i = 0$
- $cov(x, y) = y'Dx =< x, y >_D = \sum_i p_i x_i y_i$
- $var(x) = x'Dx = \|x\|_D^2$ and $\sigma(x) = \|x\|_D$
- $r(x, y) = \dfrac{cov(x, y)}{\sigma(x)\sigma(y)} = \cos(x, y)$
- $\mathbb{V} = X'DX$ is the matrix of covariances (correlations, with standardized variables)

$$\|X\|_D^2 = trace(X'DX) = \sum_i var(X^i) = \text{Total Variance}$$

Orthogonal Projections

- $\Pi_x^D = xx'D/\|x\|_D^2$ the D-orthogonal projector on $span(x)$
- $\Pi_X^D = X(X'DX)^+X'D$ the D-orthogonal projector on $span(X)$
  - Idempotent, $(\Pi_X^D)^2 = \Pi_X^D$
  - D-symmetrical matrix, $D\Pi_X^D = (\Pi_X^D)'D$
  - $\hat{y} = \Pi_X^D y = X\hat{\beta}$

    where $\hat{\beta} = (X'DX)^+X'Dy$ is the OLS estimator
  - $R^2(y; span(X)) = \frac{var(\hat{y})}{var(y)} = r^2(\hat{y}, y)$
  - $\Pi_X^D = \sum_i \Pi_{X^i}^D$ when $\{X^i\}_i$ are mutually uncorrelated

from the row's side

- $(I\!R^p, M)$ the Euclidian space for rows
- M, $p \times p$, symmetric positive definite matrix

Because variables are centered, the mean point of the cloud of the weighted rows, $\mathcal{N} = \{(X_i, p_i)\}_{i=1,n}$, is at the origin of coordinates

$$1'DX = [\overline{X}^1, ..., \overline{X}^p] = 0_{I\!R^p}\,.$$

Denote

$$\mathbb{W} = XMX' = [W_i^j = X_i M X_j{}']$$

the $n \times n$ matrix of the M-scalar products between rows.

The inertia of $\mathcal{N}$ with respect to the mean point is

$$I = \sum_{i=1}^{n} p_i \|X_i\|_M^2 = trace(\mathbb{W}D) = trace(\mathbb{V}M)\,.$$

Using the vec operator that transforms a matrix into a vector by stacking the columns one underneath the other, it can be shown, [3, Durand], that

$$trace(XMX'D) = vec'(X')(D \otimes M)vec(X') = \|vec(X')\|^2_{D\otimes M}$$

where $D \otimes M$ is the kronecker product (see [8, Magnus & Neudecker]) of the two metrics which is also a metric on $\mathbb{R}^{np}$.

Identifying $(\mathbb{R}^{n\times p}, M, D)$ and $(\mathbb{R}^{np}, D \otimes M)$,

a data set to be analyzed through a method that needs adapted metrics, is defined by the triple

$$(X, M, D)$$

and the associated squared Hilbert-Schmidt norm

$$\|X\|^2_{HS} = trace(XMX'D)$$

can be interpreted as the inertia of the row-points with respect to their centroid.

2. PCA and SVD of $(X_{n\times p}, M_{p\times p}, D_{n\times n})$

- $p \times p$ Operator for columns : $\mathbb{V}M = X'DXM$
- $n \times n$ Operator for rows : $\mathbb{W}D = XMX'D$
- Inertia : $\|X\|^2_{HS} = trace(XMX'D) = trace(X'DXM)$

Theorem 1: The SVD of (X, M, D)

Let $r = rank(X)$, there exists

- $U_{n\times r}$, D-orthonormal $(U'DU = I_r)$ whose column U^i is the eigenvector associated to the (positive) eigenvalue λ_i of $\mathbb{W}D$,
- $V_{p\times r}$, M-orthonormal $(V'MV = I_r)$ whose columns V^i is the eigenvector associated to the same eigenvalue λ_i of $\mathbb{V}M$,
- $\Lambda_r^{1/2} = diag(\lambda_1^{1/2}, ..., \lambda_r^{1/2})$ diagonal matrix of the singular values of (X, M, D),

such that

$$X = U\Lambda_r^{1/2}V' = \sum_{i=1}^{r} \sqrt{\lambda_i} U^i V^{i\prime}.$$

$$I = \|X\|^2_{HS} = trace(\Lambda_r) = \sum_{i=1}^{r} \lambda_i \, .$$

Note the ordering (the reason is explained below)

$$\lambda_1 \geq ... \geq \lambda_r > 0$$

that induces an order for the columns of U and V.

Transition Formulae and Row-Column Duality

$$U = X \, MV \Lambda_r^{-1/2}$$

$$V = X' DU \Lambda_r^{-1/2}$$

Duality	Columns	$\longleftrightarrow$	Rows
Linear Space	$span(X) \subset I\!R^n$	$\longleftrightarrow$	$span(X') \subset I\!R^p$
Metric	D	$\longleftrightarrow$	M
Matrix	X	$\longleftrightarrow$	X'
$\perp$ factorial axes	$U = [U^1 ... U^r]$	$\longleftrightarrow$	$V = [V^1 ... V^r]$

The Singular Value Decomposition of (X, M, D) provides two auxiliary decompositions

$$\mathbb{V} = V \Lambda_r V' = \sum_i \lambda_i V^i V^{i\prime} \text{ and } \mathbb{W} = U \Lambda_r U' = \sum_i \lambda_i U^i U^{i\prime}.$$

Theorem 2: Eckart-Young approximation

Let k an integer value such that $1 \leq k \leq r = rank(X)$, and $E_k = \{Z \in I\!R^{n \times p} \,|\, rank(Z) = k\}$.
Denote
$U_k = [U^1 ... U^k]$, $V_k = [U^1 ... V^k]$, $\Lambda_k^{1/2} = diag(\lambda_1^{1/2}, ..., \lambda_k^{1/2})$,
then,

$$\min_{Z \in E_k} \|X - Z\|^2_{HS} = \|X - \hat{X}_k\|^2_{HS} = \sum_{i=k+1}^{r} \lambda_i,$$

the optimum being reached by the incomplete SVD of rank k,

$$\hat{X}_k = U_k \Lambda_k^{1/2} V_k{}'.$$

The eigenvalues are decreasingly ordered $\Rightarrow$ the error is minimum.

$$\hat{X}_k = \Pi^D_{U_k} X \quad \text{and} \quad \hat{X}_k{}' = \Pi^M_{V_k} X'.$$

PCA of order k of (X, M, P)

The PCA of order k, $k \leq r$, is the incomplete SVD of rank k

$$\hat{X}_k = U_k \Lambda_k^{1/2} V_k{}'$$

as defined in the Eckart-Young theorem.

Theorem 3: PCA and the Inertia Principle (row point of view)

To look at row-points, the $\{V^i\}$ family provides optimal successive orthogonal projections,

$$\Pi^M_{V^i} X' = V^i V^{i\prime} M X', \text{ for } i = 1, ..., k,$$

according to the Inertia Principle.

If one accepts as a principle, that the best one dimensional orthogonal projection of row-points is that of maximal inertia, then V^1 is the best axis; V^2 is the best second, orthogonal to the first...

Notice that the inertia of the projected row-points whose coordinates on V^i are given by $C^i = XMV^i$, is expressed as

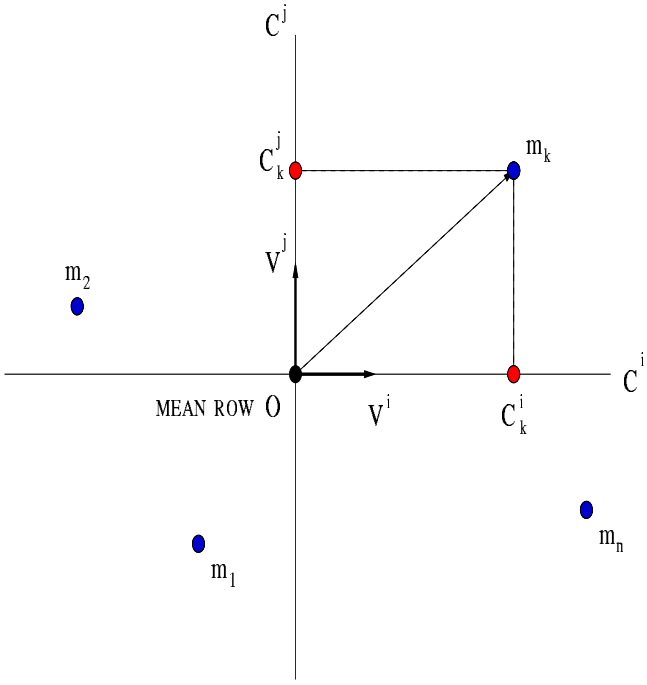
$$V^{i\prime} M \mathbb{V} M V^i = C^{i\prime} D C^i = var(C^i).$$

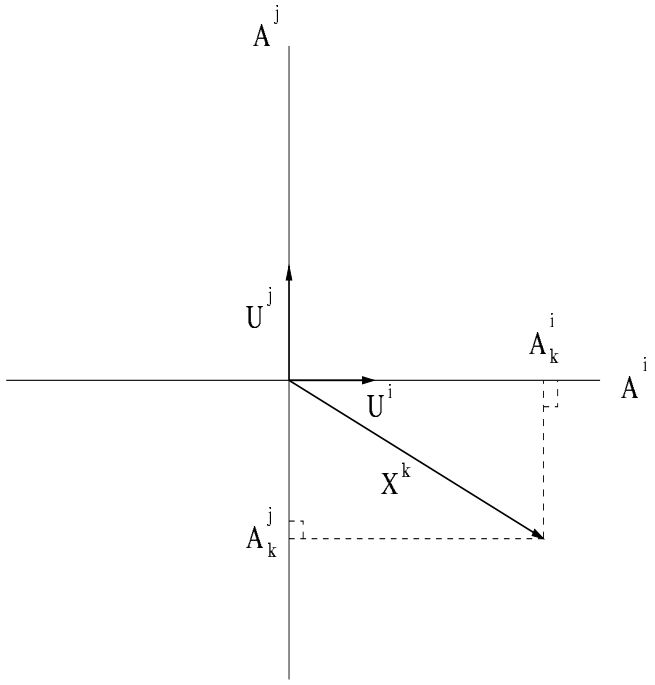
This expression may be also considered as the variance of a **Latent Variable** called the **Principal Component**

$$C^i = XMV^i = \sqrt{\lambda_i} U^i,$$

which is a linear compromise of the column-variables.

Dual plots	Rows	$\longleftrightarrow$	Columns
factorial axis	V^i	$\longleftrightarrow$	U^i
Coordinates	$C^i = \sqrt{\lambda_i} U^i$	$\longleftrightarrow$	$A^i = \sqrt{\lambda_i} V^i$
Quality of point l	$\cos^2 \theta_l^i = \dfrac{(C_l^i)^2}{\sum_j (C_l^j)^2}$	$\longleftrightarrow$	$\cos^2 \theta_l^i = \dfrac{(A_l^i)^2}{\sum_j (A_l^j)^2}$





PCA ≡ Partial Regressions on components.

(column dual version of Theorem 3)

deflated X = residuals of partial regressions

Denote $X_{(0)} = X$, the deflated X matrix at step $j \in \{1, .., k\}$ is

$$X_{(j)} = X_{(j-1)} - \Pi^D_{Uj} X_{(j-1)}$$

associated to the Partial Least-Square Regression

$$\hat{X}_{(j)} \doteq \Pi^D_{Uj} X_{(j-1)} = \Pi^D_{Uj} X = \sqrt{\lambda_j} U^j V^{j\prime} = C^j V^{j\prime}.$$

$$X = \sum_{j=1}^{k} \hat{X}_{(j)} + X_{(k)} = U_k \Lambda_k^{1/2} V_k{}' + X_{(k)}$$

The deflated covariance matrix is

$$\mathbb{V}_{(j)} = X'_{(j)} D X_{(j)} = \mathbb{V} - \sum_{i=1}^{j} \lambda_i V^i V^{i\prime}.$$

PCA	$X_{(0)} = X$	
Step j	**1) Building C^j for fixed $X_{(j-1)}$**	$C = X_{(j-1)} M V$ $V^j = arg \max_{V'MV=1} var(C)$ $C^j = X_{(j-1)} M V^j = X M V^j$
	2) Deflation	$X_{(j)} = X_{(j-1)} - \Pi^D_{Cj} X_{(j-1)}$

Alternating Least-Squares algorithm

Optimization problem 1) is solved by the largest eigenvector

$$X_{(j-1)}{}'DX_{(j-1)}MV^j = \lambda V^j$$

that suggested to [13, H. Wold] an adaptation of the "powered iterative" method through the transition formulae:

$X_{(0)} = X$

for $j = 1, ..., k$

$C^j \leftarrow \mathbf{1}_n$

repeat until convergence of V^j

$V^j \leftarrow X_{(j-1)}{}'DC^j/(C^{j\prime}DC^j)$ * regression of $X_{(j-1)}$ on C^j

$V^j \leftarrow V^j/\sqrt{V^{j\prime}MV^j}$ making the norm equal to 1

$C^j \leftarrow X_{(j-1)}MV^j/(V^{j\prime}MV^j)$ * regression of $X_{(j-1)}{}'$ on V^j

end repeat

$X_{(j)} \leftarrow X_{(j-1)} - C^jV^{j\prime}$ deflation of $X_{(j-1)}$

Notice that this algorithm is used **in presence of missing data**. The rule is:

Compute the regression coefficients in * on the remaining data.

Methods derived from adapted metrics

Method	X	M	D
usual PCA	p continuous variables	I_p	$n^{-1}I_n$
Correspondence Analysis of P (frequencies)	$D_r^{-1}PD_c^{-1}$	$D_c = diag(P'\mathbf{1})$	$D_r = diag(P\mathbf{1})$
Discriminant Analysis of n groups on $T_{N\times p}$	$G_{n\times p}$ group centroids	$\mathbb{V}^{-1}$ Mahalanobis computed on T	$diag(p_1, .., p_n)$ group weights
PCA of (Y, M, D) on Instrumental Variables T	$\hat{Y} = \Pi_T^D Y$	M	D

and others that derive from these ones...

Principal Component Regression (PCR), Multiple Correspondence Analysis (MCA), Canonical Correlation Analysis (CCA), Redundancy Analysis, Non Symmetrical Correspondence Analysis,...

3. Linear PLS regression

The context, in the usual way of presenting $PLS(X, Y)$, see [14, S. Wold et al.],

- $(X_{n\times p}, I_p, D)$ for predictors
- $(Y_{n\times q}, I_q, D)$ for responses.

Algorithm and Model

In contrast to PCR, PLS constructs components $\{t_j\}_{j=1,...,k}$ as linear compromises of X, on which Y Partial Regressions are successively processed.

PLS	$X_{(0)} = X, \quad Y_{(0)} = Y$
Step j	**1) Building** t^j **for fixed** $(X_{(j-1)}, Y_{(j-1)})$
	$t = X_{(j-1)}w,\ u = Y_{(j-1)}v$ $(w^j, v^j) = arg \max_{w'w=1=v'v} cov(t, u)$ $t^j = X_{(j-1)}w^j,\ u^j = Y_{(j-1)}v^j$
	2) Deflations
	$X_{(j)} = X_{(j-1)} - \Pi^D_{tj} X_{(j-1)}$ $Y_{(j)} = Y_{(j-1)} - \Pi^D_{tj} Y_{(j-1)}$

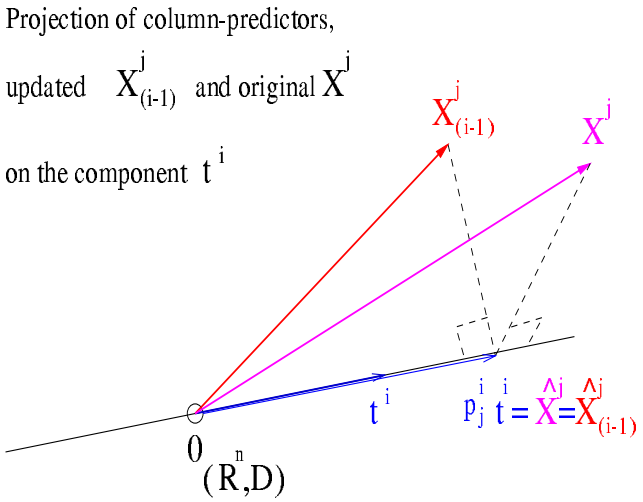
$\boldsymbol{PLS(X, Y = X) \equiv PCA(X)}$ (usual).

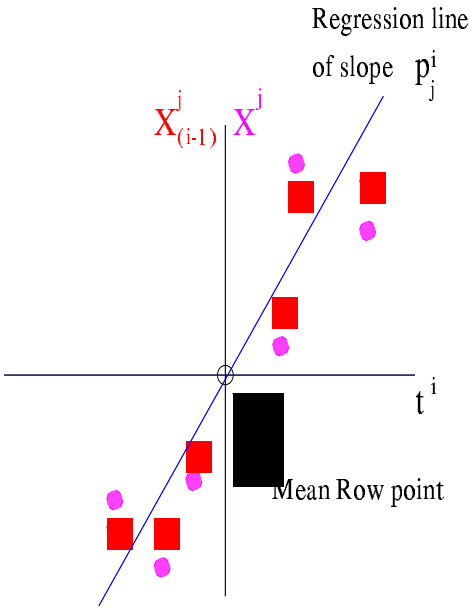
Alternating Least-Squares algorithm (NIPALS), for missing data.

Components $\{t^i\}$

- belong to $span(X)$, that is
$$t^i = X w*^i$$
- are mutually D-orthogonal
$$t^{j\prime} D t^i = 0, \quad i \neq j \ .$$

$$\hat{X}_{(i)} = \Pi^D_{t^i} X_{(i-1)} = \Pi^D_{t^i} X = t^i p^{i\prime}$$
$$\hat{Y}_{(i)} = \Pi^D_{t^i} Y_{(i-1)} = \Pi^D_{t^i} Y = t^i c^{i\prime}$$





Partial Regressions of updated variables give the same results as regressions of the original variables on the components.

Denoting

$$T(k) = [t^1 ... t^k], \quad W(k) = [w^1 ... w^k],$$

$$P(k) = [p^1 ... p^k], \quad C(k) = [c^1 ... c^k],$$

then, $W(k)'W(k) = I_k$ and $span(W(k)) \subset span(X')$.

There exists a relationship between the families $\{w*^i\}$ and $\{w^i\}$

$$W^*(k) = [w*^1 ... w*^k] = W(k)(P(k)'W(k))^{-1} .$$

The PLS fit from both X and Y sides, is

$$\hat{X}(k) = \sum_{i=1}^{k} \hat{X}_{(i)} = \Pi^D_{T(k)} X = t^1 p^{1\prime} + \ldots + t^k p^{k\prime} = T(k)P(k)' ,$$

$$\hat{Y}(k) = \sum_{i=1}^{k} \hat{Y}_{(i)} = \Pi^D_{T(k)} Y = t^1 c^{1\prime} + \ldots + t^k c^{k\prime} = T(k)C(k)' ,$$

so that, the models are given by

$$X = \hat{X}(k) + X_{(k)} , \qquad Y = \hat{Y}(k) + Y_{(k)} .$$

If $k = rank(X)$, then $span(T(k)) = span(X)$, and

$$PLS(X, Y) = OLS(X, Y),$$

thus providing an upper bound for the number of components

$$1 \leq k \leq rank(X) = r.$$

Projection based representations, the gap from duality

We aim at finding maps for column and row points that are based on orthogonal projections, [3, Durand].

Predictors	Columns	Rows
Linear Spaces	$span(X) \subset I\!R^n$	$span(X') \subset I\!R^p$
Metrics	D	I_p
$\perp$ families	$\{t^i = X_{(i-1)} w^i\}$	$\{w^i\}$
Metrics	D	$\mathbb{V} = X'DX$
$\perp$ families	$\{t^i = X w*^i\}$	$\{w*^i\}$

From the variable's side

	Predictor X^j	Response Y^j
Coordinate on $\tilde{t}^i = t^i / \|t^i\|_D$	$r(t^i, X^j)\sigma(X^j)$	$r(t^i, Y^j)\sigma(Y^j)$
Measure of quality	$r^2(t^i, X^j)$	$r^2(t^i, Y^j)$

From the row's side

The only way of representing X-row points is to use the $\mathbb{V}$-orthogonal $\{w*^i\}$ family

$$w*^j{}'\mathbb{V}w*^i = t^{j\prime}Dt^i = 0, \quad i \neq j.$$

Denote

$$\tilde{w}*^i = w*^i / \| w*^i \|_{\mathbb{V}} = w*^i / \|t^i\|_D.$$

Theorem 4 : Coordinates of projected row-points

The coordinates of the X-row points projected on the axis $\tilde{w}*^i$ are given by the vector t^{*i} defined by

$$t^{*i} = \mathbb{W}D\tilde{t}^i = \mathbb{W}Dt^i/\|t^i\|_D.$$

Proof:

$$\Pi^{\mathbb{V}}_{\tilde{w}^{*i}} X' = \tilde{w}^{*i}\tilde{w}^{*i\prime}\mathbb{V}X' = \tilde{w}^{*i}(XX'DX\tilde{w}^{*i})' \,. \,\square$$

The gap from duality

Alas! In contrast to the perfect dual case, the vector of coordinates of projected row-points, t^{*i}, and the axis on which the variables are projected, the component t^i, are not collinear (except in two cases).

For the component t^i, the gap from duality is measured by

$$GFD_i = 1 - r(t^i, t^{*i})$$

that lies within $[0, 1]$.

Two extreme cases of perfect duality

Case $GFD_i = 0$: $\exists \alpha > 0$ such that $XX'Dt^i = \alpha t^i$	
$Y = X$	$PLS(X, Y) \equiv PCA(X)$, then $\alpha = \lambda_i$
$\mathbb{V} = I_p$	non correlation, then $\alpha = 1$

The closer to 0 is GFD_i, the easier is the interpretation of the X row-points, located at t^{*i}, by the predictors and responses projected on t^i.

Measure of quality for the representation of the row-point l on the axis $w*^i$

$$\cos^2 \theta_l^i = \frac{(t_l^{*i})^2}{\sum_{j=1}^r (t_l^{*j})^2}.$$

The building-model stage: choosing the dimension k, $1 \leq k \leq r$

- Based on fit criteria: (X, Y) adjusted by $(\hat{X}(k), \hat{Y}(k))$
  - From the X side:
    - Proportion of the reconstructed variance

$$\mathcal{I}^x(k) = \sum_{l=1}^{p} var(X^l) R^2(X^l; Im\, T(k)) / trace(X'DX)$$

$$\mathcal{I}^x(r) = 1.$$

    - Proportion of the reconstructed $\mathbb{V}$-inertia of X-row points

$$\mathcal{I}^{ind}(k) = \sum_{i=1}^{k} var(t^{*i}) / trace(\mathbb{V}^2)$$

$$\mathcal{I}^{ind}(r) = 1.$$

  - From the Y side: Proportion of the reconstructed variance

$$\mathcal{I}^y(k) = \sum_{h=1}^{q} var(Y^h) R^2(Y^h; Im\, T(k)) / trace(Y'DY)$$

$$\mathcal{I}^y(r) \leq 1.$$

- Based on internal prediction criterion: (leave-one-out) C.V.

$$PRESS(k) = \sum_{j=1}^{q} \frac{1}{n} \sum_{i=1}^{n} (Y_i^j - X_i \hat{\beta}_{(-i)}|^j)^2 .$$

PLS as a unifying framework for D. A.

The context, in the general way of presenting $PLS(X, M_x; Y, M_y)$,

- $(X_{n\times p}, M_x, D)$ for predictors
- $(Y_{n\times q}, M_y, D)$ for responses.

$$PLS(X, M_x\,; Y = X, M_y = M_x) \equiv PCA(X, M_x, D),$$

$\Rightarrow$ perspectives to non symmetrical multivariate methods.

Linear compromises in the PLS algorithm are now

$$t^i = X_{(i-1)} M_x w^i \quad \text{and} \quad u^i = Y_{(i-1)} M_y v^i$$

Predictors	Columns	Rows
Linear Spaces	$span(X) \subset I\!R^n$	$span(X') \subset I\!R^p$
Metrics $\perp$ families	D $\{t^i = X_{(i-1)} M_x w^i\}$	M_x $\{w^i\}$
Metrics $\perp$ families	D $\{t^i = X M_x w*^i\}$	$M_x \mathbb{V} M_x = M_x X' D X M_x$ $\{w*^i\}$

The $M_x \mathbb{V} M_x$-inertia of row-points becomes $I^{ind} = trace((\mathbb{V} M_x)^2)$.

The coordinates of projected row-points on $w*^i$ are now

$$t*^i = X M_x X' D t^i / \|t^i\|_D \,.$$

4. Splus examples

First example: the Cornell data

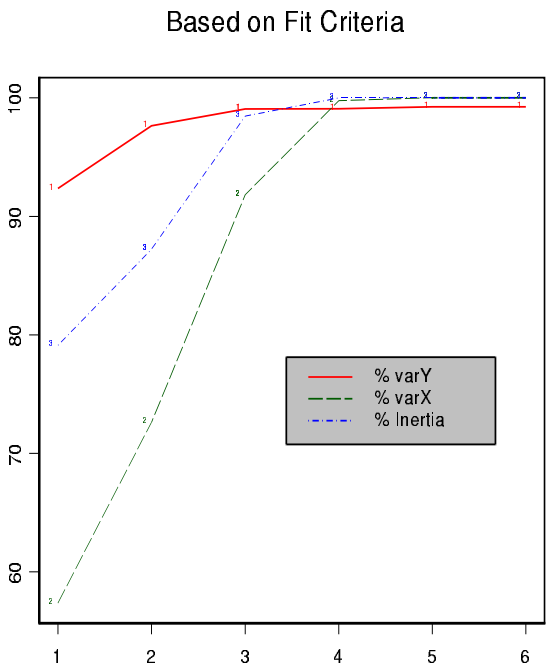
The response y is the octane rating of 12 different mixtures to determine the influence of 7 constituents, see [12, Tenenhaus]. The 7 predictors represent proportions and sum to 1.

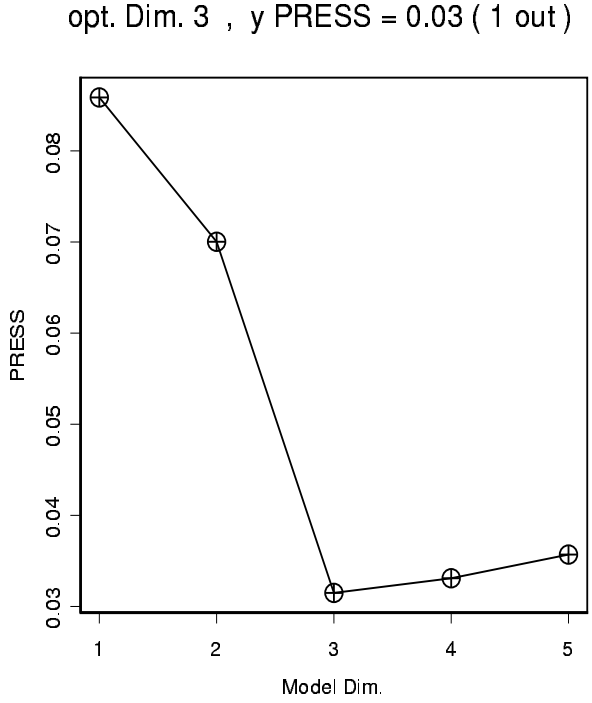
n^o	x1	x2	x3	x4	x5	x6	x7	y
1	0.00	0.23	0.00	0.00	0.00	0.74	0.03	98.7
2	0.00	0.10	0.00	0.00	0.12	0.74	0.04	97.8
3	0.00	0.00	0.00	0.10	0.12	0.74	0.04	96.6
4	0.00	0.49	0.00	0.00	0.12	0.37	0.02	92.0
5	0.00	0.00	0.00	0.62	0.12	0.18	0.08	86.6
6	0.00	0.62	0.00	0.00	0.00	0.37	0.01	91.2
7	0.17	0.27	0.10	0.38	0.00	0.00	0.08	81.9
8	0.17	0.19	0.10	0.38	0.02	0.06	0.08	83.1
9	0.17	0.21	0.10	0.38	0.00	0.06	0.08	82.4
10	0.17	0.15	0.10	0.38	0.02	0.10	0.08	83.2
11	0.21	0.36	0.12	0.25	0.00	0.00	0.06	81.4
12	0.00	0.00	0.00	0.55	0.00	0.37	0.08	88.1
moy.	0.07	0.22	0.04	0.25	0.04	0.31	0.06	88.58
stdev	0.09	0.19	0.05	0.22	0.05	0.28	0.03	6.24

The matrix of correlations shows that the predictors x_1 and x_3 are strongly correlated as well as x_4 and x_7. The response y presents a strong correlation with x_6 and at a lesser degree with x_1 and x_3.

The predictor x_5 is poorly correlated with the other variables.

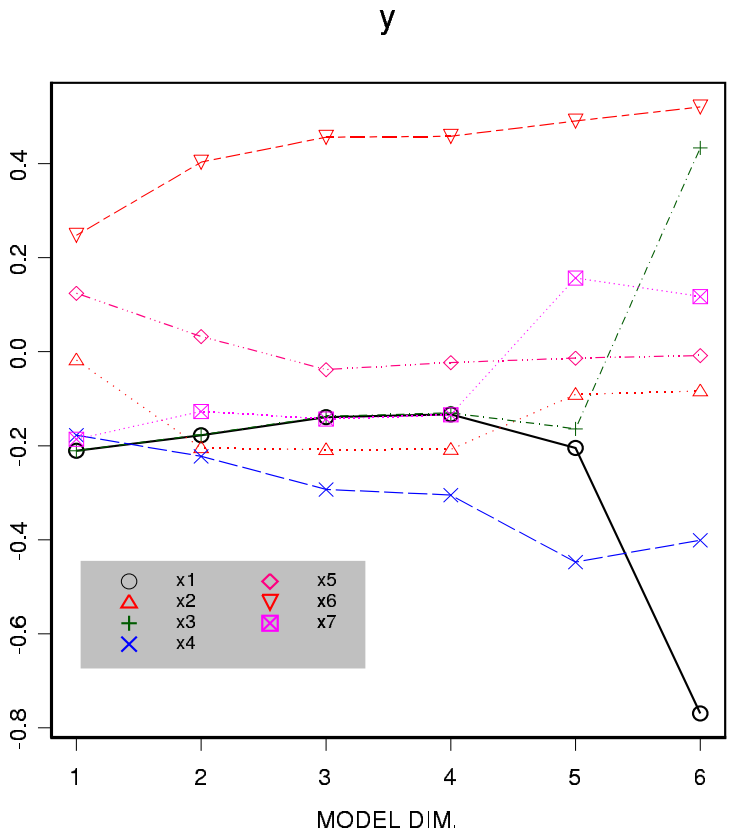
	x2	x3	x4	x5	x6	x7	y
x1	0.10	0.999	0.37	-0.55	-0.80	0.60	-0.84
x2		0.10	-0.54	-0.29	-0.19	-0.59	-0.07
x3			0.37	-0.55	-0.81	0.61	-0.84
x4				-0.21	-0.65	0.92	-0.71
x5					0.46	-0.27	0.49
x6						-0.66	0.99
x7							-0.74





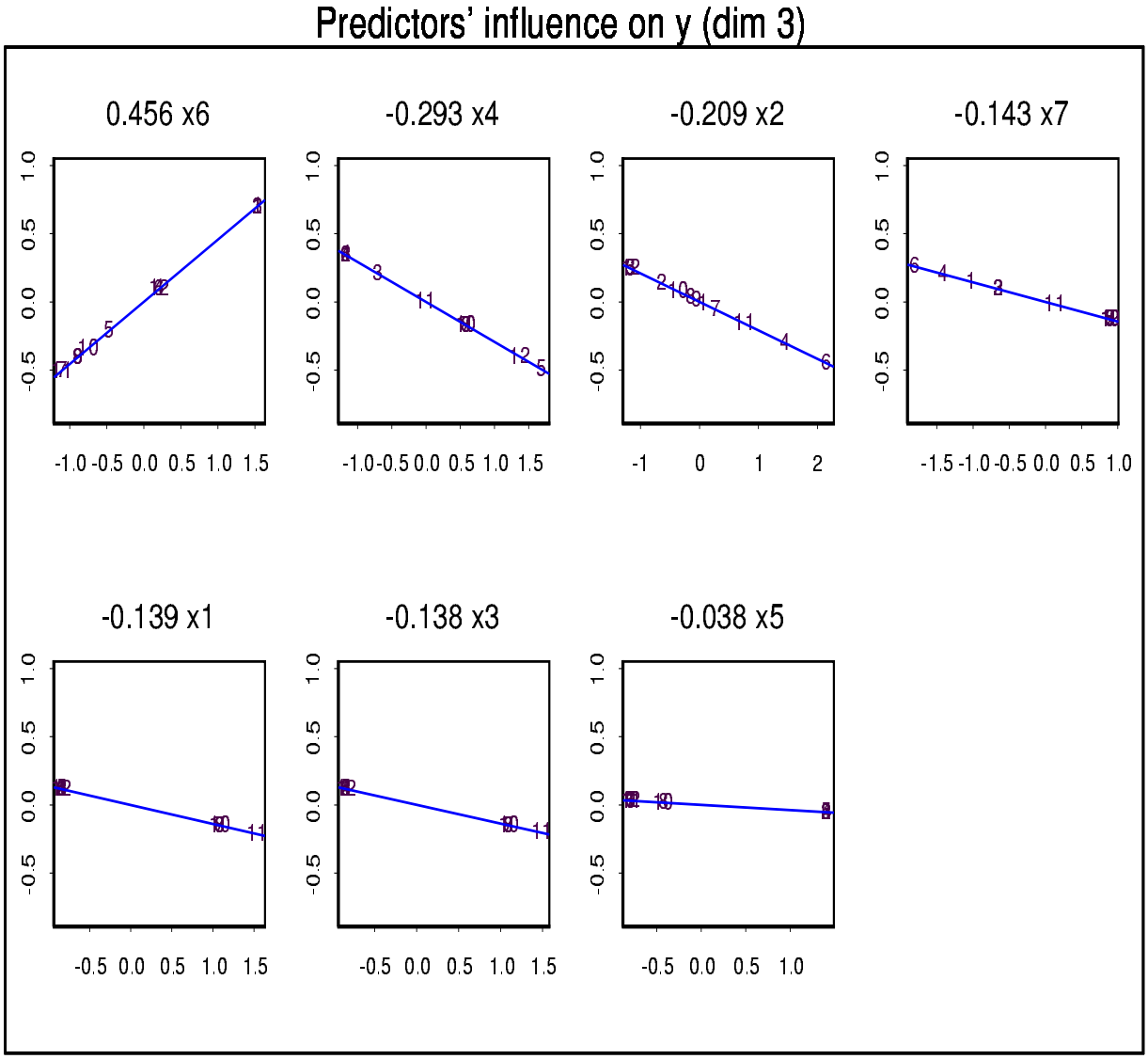
The building-model stage: $k = 3$.

The evolution of the coefficients of the standardized variables in models with successive dimensions, presents an another information to choose the best dimension.



Evolution of coefficients $\hat{\beta}(k)$ along with the dimension k.

Finally, the retained dimension is $k = 3$ and the coordinate function plots, ordered from left to right and top to bottom in decreasing order of the coefficients, show the additive influence of the observations on the fit $\hat{y}(3)$.

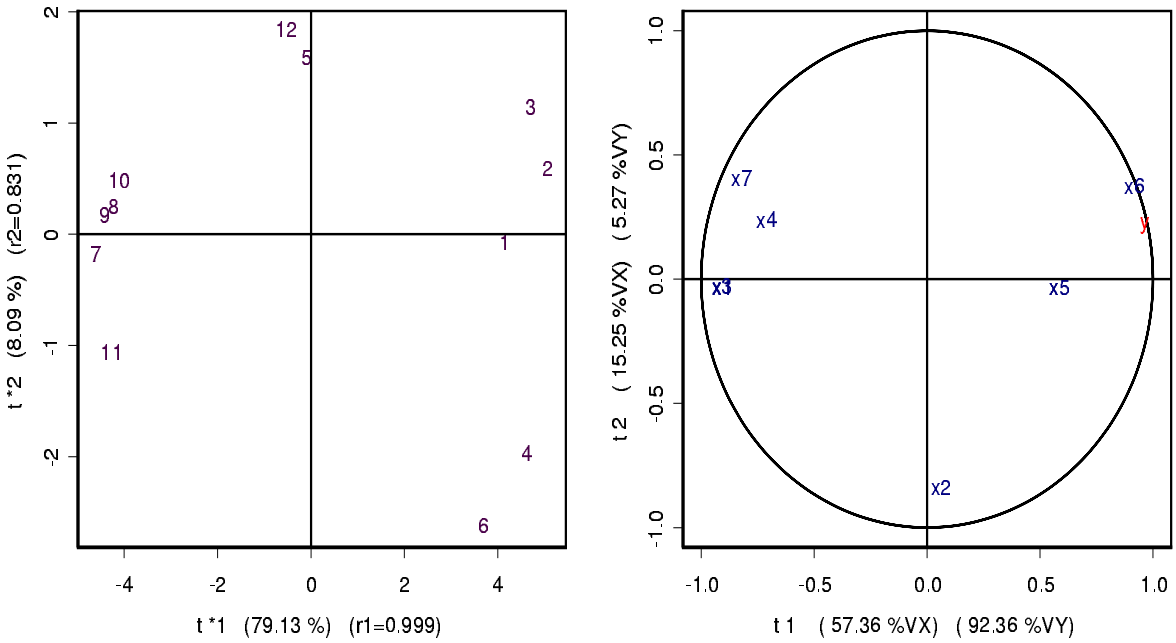


Observations and coordinate functions $(x^i, \hat{\beta}_i(3)x^i)$.

Projection based representations

Are indicated:

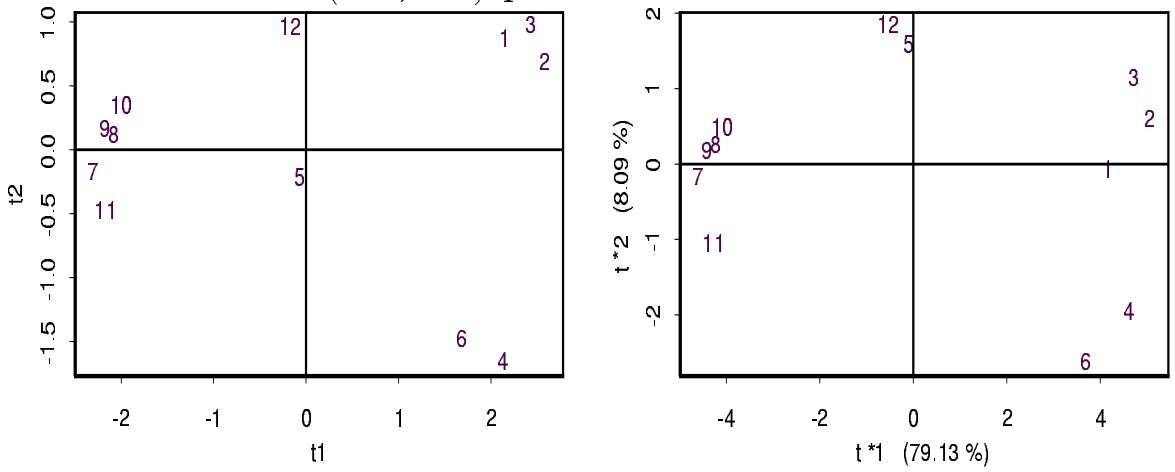
- for the observations, the percentage of the contribution of the axis to the inertia
- for the variables, the percentage of the contribution of the axis to the variance.



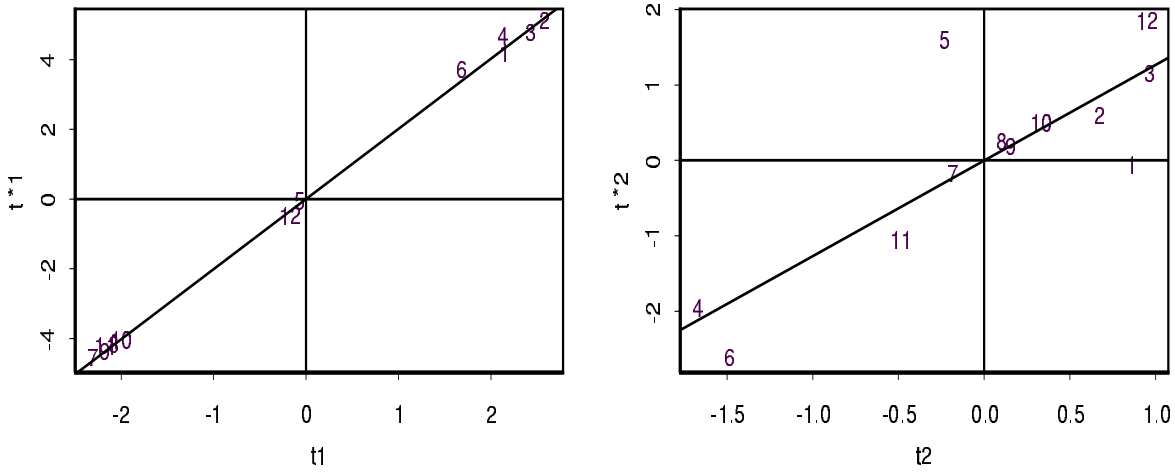
(1,2) scatter plots for observations and variables, the gaps from duality are very small, they correspond to $r_1 = 0.999$ *et* $r_2 = 0.831$.

GFD_1	GFD_2	GFD_3	GFD_4	GFD_5	GFD_6
0.001	0.169	0.112	0.008	0.651	0.016
0.999	0.831	0.888	0.992	0.349	0.984
r_1	r_2	r_3	r_4	r_5	r_6

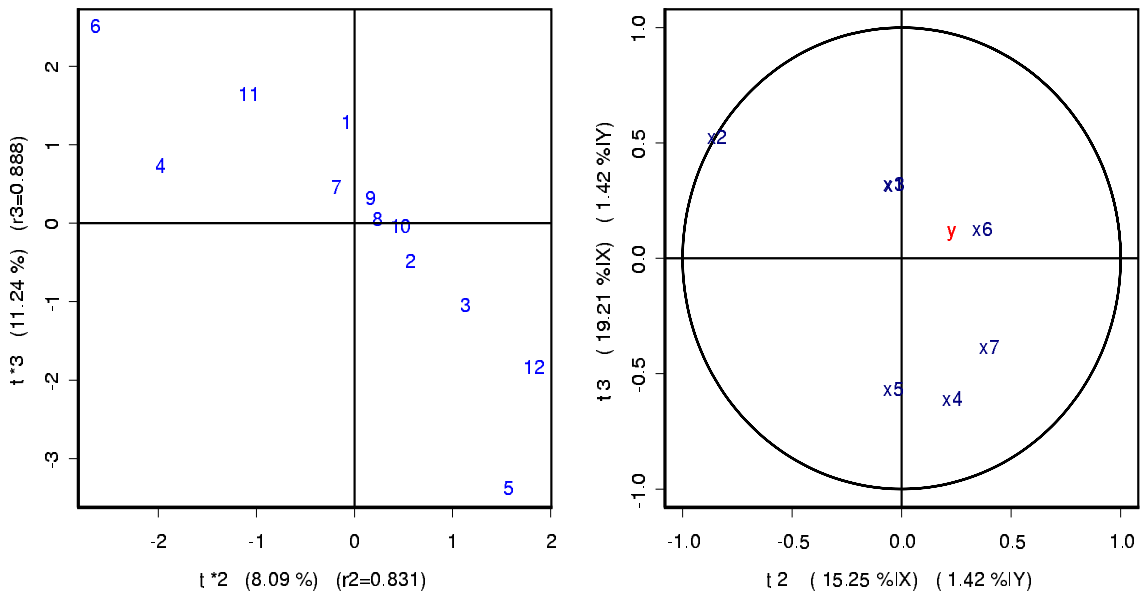
Comparison between pseudo-factorial (t^1, t^2) and factorial $(t*^1, t*^2)$ plots of observations.



r1 = 0.999 r2 = 0.831



	$\cos_1^2$	$\cos_2^2$	$\cos_3^2$	$\cos_4^2$	$\cos_5^2$	$\cos_6^2$
1	0.8951	0.0004	0.0837	0.0207	0.0001	0
2	0.9601	0.0123	0.0094	0.0182	0.0001	0
3	0.8868	0.0511	0.0453	0.0165	0.0002	0
4	0.8211	0.1511	0.0200	0.0078	0.0000	0
5	0.0007	0.1768	0.8224	0.0001	0.0000	0
6	0.4946	0.2524	0.2250	0.0279	0.0001	0
7	0.9893	0.0017	0.0089	0.0000	0.0000	0
8	0.9937	0.0031	0.0001	0.0032	0.0000	0
9	0.9942	0.0013	0.0045	0.0000	0.0000	0
10	0.9828	0.0129	0.0002	0.0040	0.0000	0
11	0.8245	0.0521	0.1190	0.0043	0.0001	0
12	0.0341	0.4020	0.4106	0.1520	0.0013	0



Factorial (2,3) scatter plots that allow to "see" 5 and 12.

Second example: calibration of moisture of cheeses through near infrared spectroscopy

We aim at predict the moisture content of 56 cheeses by their absorbance measured at 150 successive wavelengths of near infrared spectra, see [9, Mazerolles et al.].

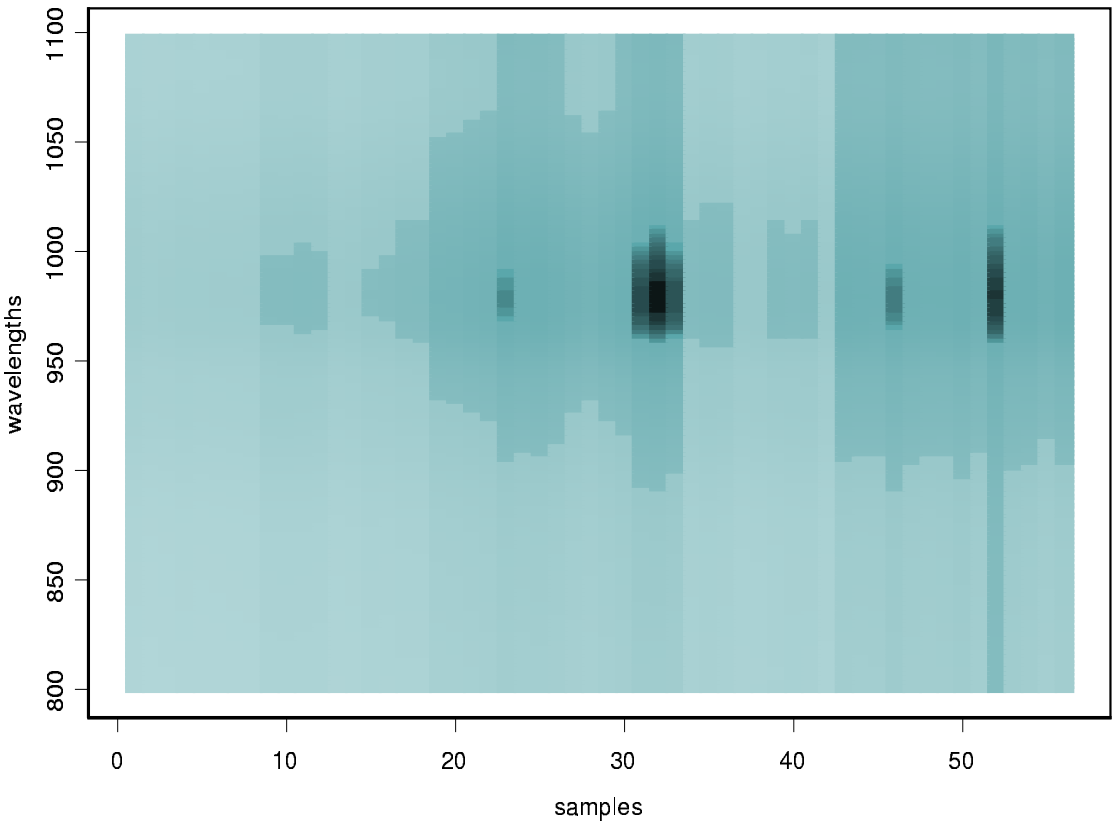
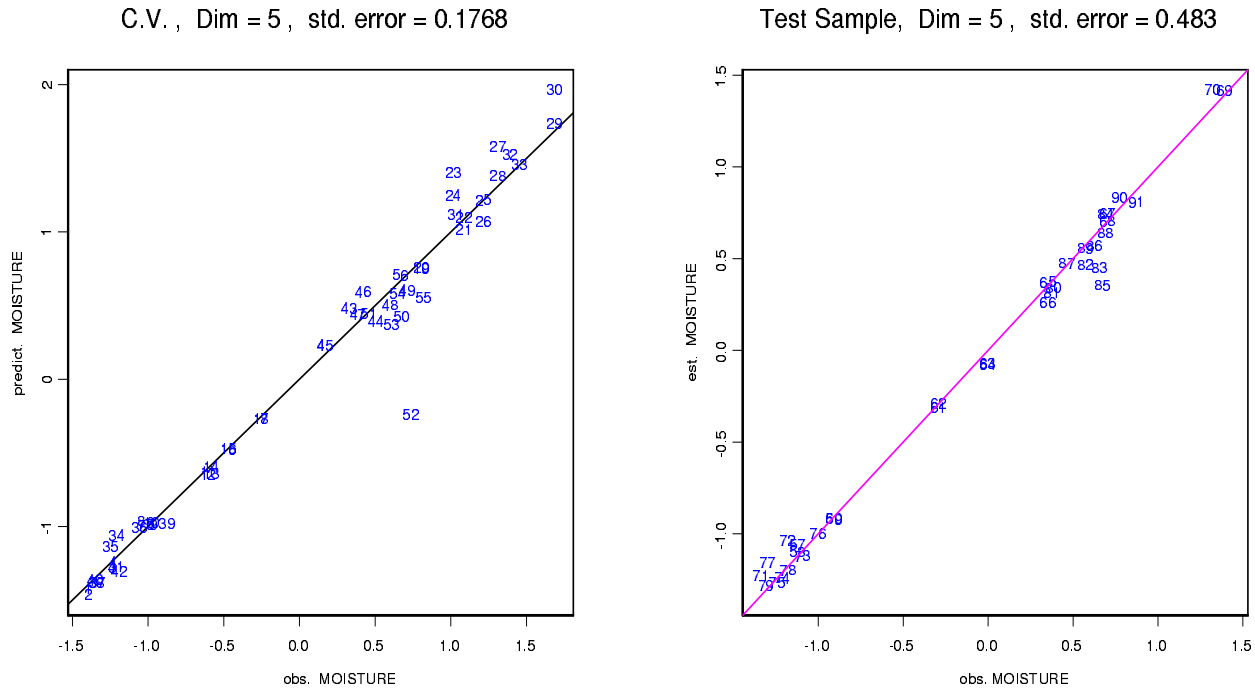


Image of the X matrix: 56 samples × 150 wavelengths.

Moreover, we dispose of a test sample of measures on 35 different cheeses.

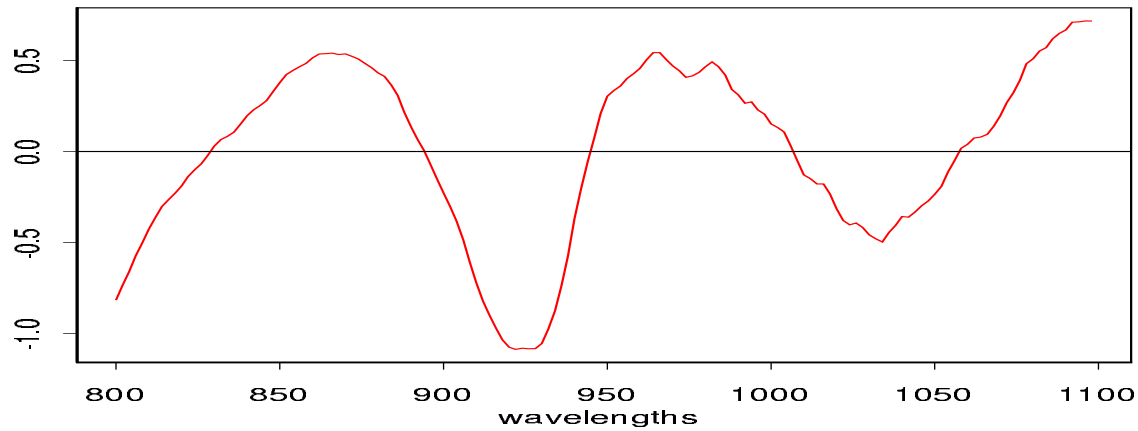


Observed $MOISTURE$ versus predicted $MOISTURE$ by Cross-Validation, left-side, and on the test sample, right-side.

The retained dimension is 5. Notice that the sample 52, badly predicted by the cross-validation, has its spectrum situated above the others.

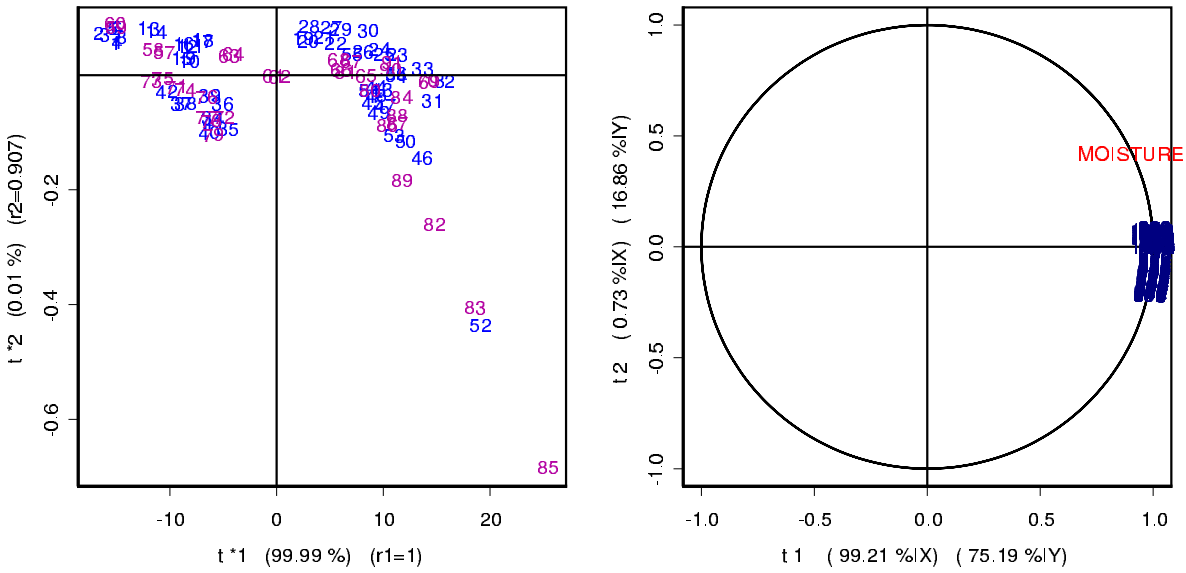
The aim of PLS calibration is also to detect influential predictors (wavelengths).

One has to look at the $\hat{\beta}(5)$ values associated to the wavelengths. The next figure shows that the regions near 925, 970 and 1020 are used by the model to predict $MOISTURE$.



Evolution of the $\hat{\beta}(5)$'s along with the wavelengths.

The sign of the coefficients: the characteristic region of water, positive near 970, cannot be dissociated from that of dry mater, negative near 925 for the fat and near 1020 for the protein.



(1,2) scatter plots for observations and variables.

Section III.

A Nonlinear Fuzzy Coding Approach through Regression Splines

1. What are regression splines?

To transform a continuous variable x whose values range within $[a, b]$, a spline function s is made of adjacent polynomials of degree d that join end to end at points called "the knots", [1, De Boor].

The linear space $S(m, \{\tau_{m+1}, \ldots, \tau_{m+K}\}, [a, b])$ of dimension $m + K$, of spline functions is characterized by three tuning parameters

- the degree d or the order $m = d + 1$ of the polynomials,
- the number K and
- the location of knots $\{\tau_{m+1}, \ldots, \tau_{m+K}\}$

$$\tau_1 = \ldots = \tau_m = a < \tau_{m+1} \leq \ldots \leq \tau_{m+K} < b = \tau_{m+K+1} = \ldots = \tau_{2m+K}.$$

A spline $s \in S(m, \{\tau_{m+1}, \ldots, \tau_{m+K}\}, [a, b])$ can be written

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

where $\{B_i^m(.)\}_{i=1,\ldots,m+K}$ is a basis of spline functions.
The vector β of the coordinate values is to be estimated by a regression method.

Notice that, when $K = 0$, $S(m, \emptyset, [a, b])$ is the set polynomials of order m on $[a, b]$.

Two sets of basis functions

- **The truncated power functions**

  The function,

  $$x \longrightarrow (x-\tau)_+^d$$

  defined by $(x-\tau)^d$ if $x \geq \tau$, 0 otherwise, is called the truncated power of degree d.

  When knots are distinct, a basis of $S(m, \{\tau_{m+1}, \ldots, \tau_{m+K}\}, [a, b])$ is given by

  $$1, x, \ldots, x^d, (x-\tau_{m+1})_+^d, \ldots, (x-\tau_{m+K})_+^d$$

- **The B-splines**

B-splines of degree d (order $m = d+1$): for $j = 1, \ldots, m+K$,

$$B_j^m(x) = (-1)^m(\tau_{j+m} - \tau_j)[\tau_j, \ldots, \tau_{j+m}](x-\tau)_+^d$$

where $[\tau_j, \ldots, \tau_{j+m}](x-\tau)_+^d$ is the divided difference of order m computed at $\tau_j, \ldots, \tau_{j+m}$ for the function $\tau \longrightarrow (x-\tau)_+^d$.

This basis is the most popular partly due to the next property that allows to compute recursively the values of B-splines, [1, De Boor]

$B_j^1(x) = 1$ if $\tau_j \leq x \leq \tau_{j+1}$, 0 otherwise,
For $k = 2, \ldots, m$,

$$B_j^k(x) = \frac{x-\tau_j}{\tau_{j+k-1}-\tau_j} B_j^{k-1}(x) + \frac{\tau_{j+k}-x}{\tau_{j+k}-\tau_{j+1}} B_{j+1}^{k-1}(x).$$

The attractive B-splines family for coding

- **Local support**:

  $$B_i^m(x) = 0, \quad \forall x \notin [\tau_i, \tau_{i+m}].$$

  ♡ → One observation x_i, has a local influence on $s(x_i)$ that depends only on the m basis functions whose supports encompass this data.

  ♠ → The counterpart is that $s(x) = 0$ outside $[a, b]$.

- **Fuzzy coding functions**:

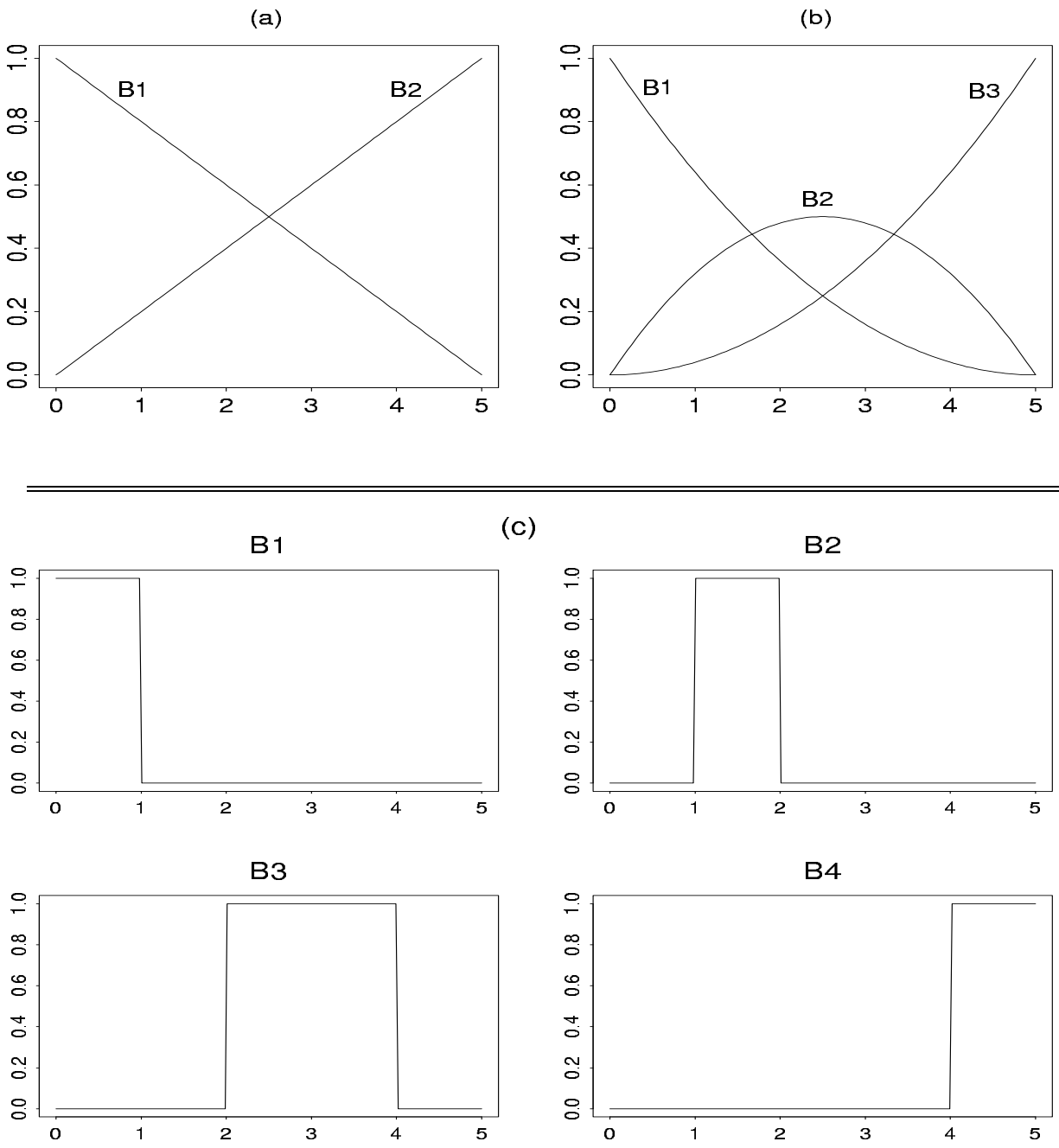
  $$0 \leq B_i^m(x) \leq 1 \ \textbf{(1)} \text{ and } \sum_{i=1}^{m+K} B_i^m(x) = 1 \ \textbf{(2)}.$$

  ♡ → $B_i^m(x)$ measures the degree of membership of x to $[\tau_i, \tau_{i+m}]$. The set $\{[\tau_i, \tau_{i+m}] \,|i = 1, ..., m+K\}$ is a fuzzy partition of $[a, b]$.

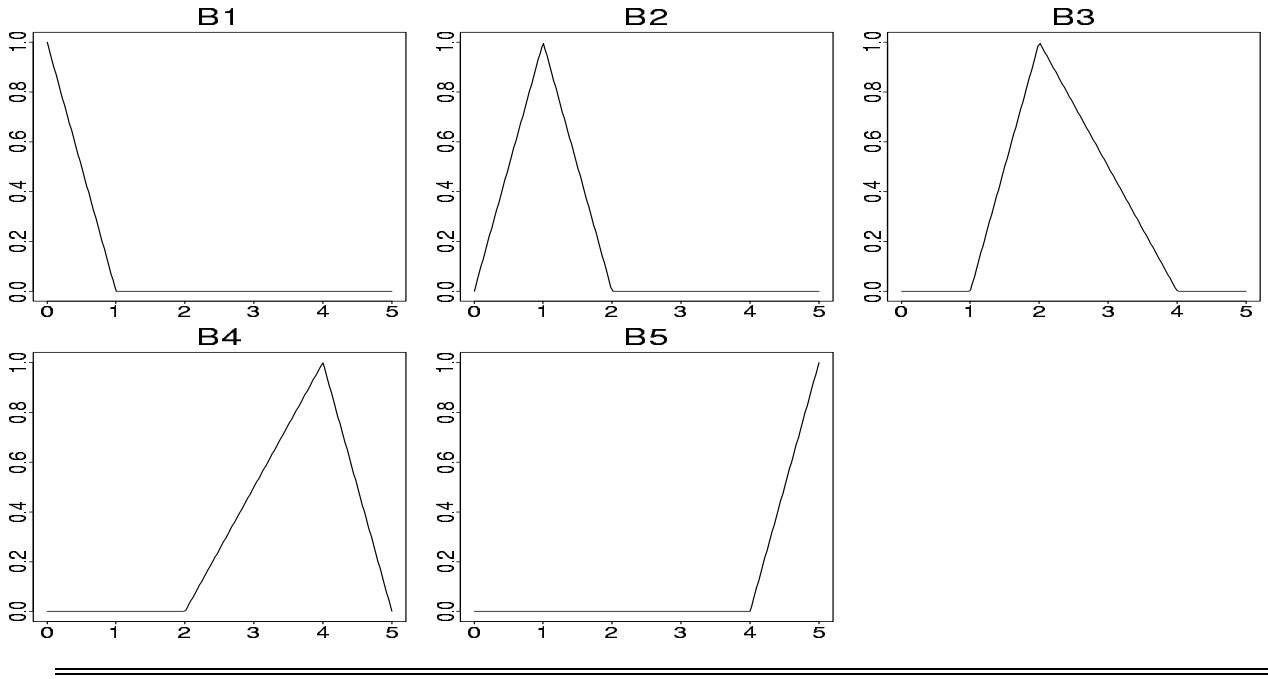
- **The multiplicity of knots controls the smoothness**:
  The multiplicity of a knot is the number of knots that merge at the same point. The multiplicity may vary from 1, a simple knot, to m, a multiple knot of order m.

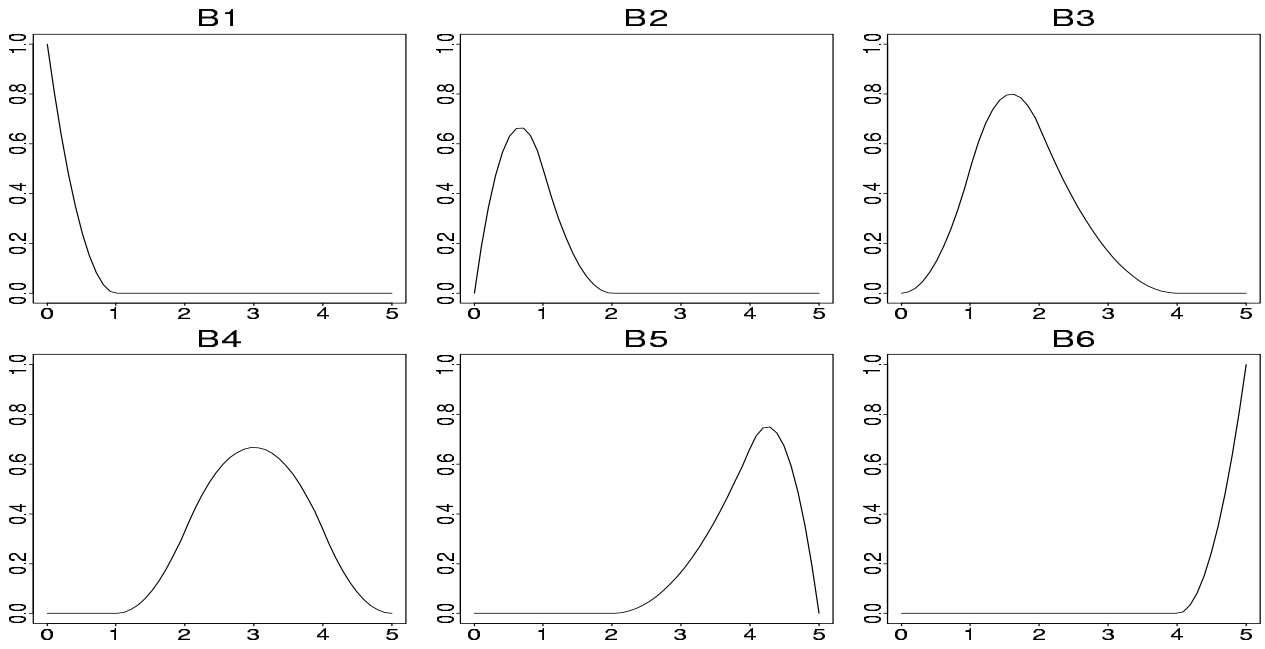
  Let m_i be the multiplicity of τ_i, $0 \leq m_i \leq m$,
  then the first $m - 1 - m_i$ right and left derivatives are equal,
  $$s_-^{(j)}(\tau_i) = s_+^{(j)}(\tau_i), \quad j \leq m - 1 - m_i.$$

  $$m_i = m \quad \Rightarrow \text{discontinuity at } \tau_i$$
  $$m_i = 1 \quad \Rightarrow \text{locally } C^{m-2} \text{ at } \tau_i.$$



B-splines for $S(2, \emptyset, [0,5])$ *(a),* $S(3, \emptyset, [0,5])$ *(b) and* $S(1, \{1,2,4\}, [0,5])$ *(c).*





B-splines for $S(2, \{1, 2, 4\}, [0, 5])$ *and* $S(3, \{1, 2, 4\}, [0, 5])$.

- **Coding a variable x through B-splines**

  Due to **(2)**, two B-splines bases are generally used:

  $\{B_j^m(x) \,|\, j = 1, \ldots, m+K\}$ usual basis

  $\{1, B_j^m(x) \,|\, j = 2, \ldots, m+K\}$, modified basis.

Let $X = (x_1, \ldots, x_n)'$ be a n-sample of the variable x, denote

$$B = [B^1(X) \ldots B^{m+K}(X)] \quad \text{or} \quad B = [B^2(X) \ldots B^{m+K}(X)]$$

the complete $n \times (m+K)$, or incomplete $n \times (d+K)$, coding matrix of the sample.

Notice that $d = 0$ provides a binary coding matrix B.

D-centering the coding matrix

When the columns of B are centered, then,

$$rank(B) \leq \min(n-1, d+K).$$

Bivariate regression splines for (x, z)

$\{1, B_1^j(x) \,|\, j \in I_1\}$ and $\{1, B_2^j(z) \,|\, j \in I_2\}$ univariate bases

$$s(x,z) \in span[\{1, B_1^j(x) | j \in I_1\} \bigotimes \{1, B_2^j(z) | j \in I_2\}]$$

$$s(x,z) = \beta_1 + \sum_{j \in I_1} \beta_j^1 B_1^j(x) + \sum_{j \in I_2} \beta_j^2 B_2^j(z) + \sum_{i \in I_1} \sum_{j \in I_2} \beta_{i,j}^{1,2} B_1^i(x) B_2^j(z)$$

is a bivariate regression spline split into the ANOVA decomposition

main effects s^1 and s^2 in x and z

$$s^1(x) = \sum_{j \in I_1} \beta_j^1 B_1^j(x) \qquad s^2(z) = \sum_{j \in I_2} \beta_j^2 B_2^j(z)$$

interaction part s^{12}

$$s^{12}(x,z) = \sum_{i \in I_1} \sum_{j \in I_2} \beta_{i,j}^{1,2} B_1^i(x) B_2^j(z)$$

$B = [B_1 | B_2 | B_{1,2}]$ column centered coding matrix from X and Z.

The Least-Squares Splines (LSS) [11, Stone]

Denote X, $n \times p$, and Y, $n \times q$, the standardized sample matrices for the p predictors and the q responses.

The centered coding matrix of X, $B = [B_1|\ldots|B_p]$, leads to q separate additive spline models, $j = 1, \ldots, q$,

$$\widehat{y}^j = s^{j,1}(x^1) + \ldots + s^{j,p}(x^p)$$

through

$$\mathbf{LSS}(X, Y) \equiv \mathbf{OLS}(B, Y) \Longleftrightarrow \widehat{Y} = \Pi_B^D Y = B\widehat{\beta} = \textstyle\sum_i B_i \widehat{\beta}_i$$

where $\widehat{\beta}_i$ is the row-block associated to the coding matrix B_i.

Tuning parameters: The spline spaces used for the predictors

LSS drawbacks: numerical instability of $(B'DB)^{-1}$ if it exists !

♠ needs a large ratio observations/(column dimension of B)

♠* needs locating knots in non empty regions

♠ perturbing concurvity effects due to correlated predictors

♠* no interaction terms involved

∘ MARS [6, Friedman] proposes to remedy ♠*: automatic ↗ ↘ procedure to select knots and high order interactions through linear truncated power functions.
∘ EBOK [10, Molinari et al.]: K fixed, optimal location of knots.
∘ others....

2. PLS through Splines (PLSS): exploring and modelling data nonlinearly

Multi-collinearity, nonlinearity and outliers: the orange juice data

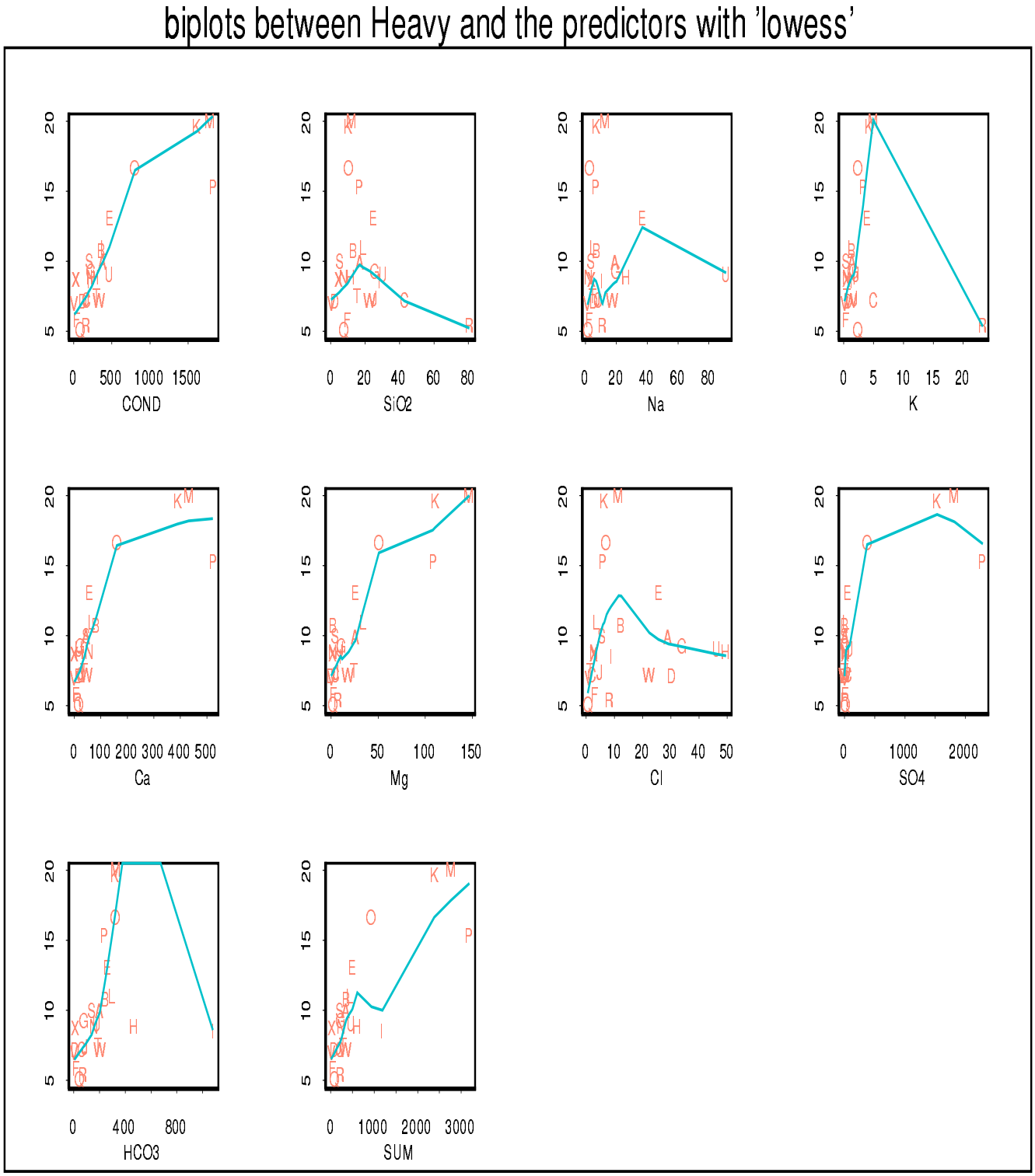
- $(x^1, \ldots, x^p)$, p predictors, $X = [X^1|\ldots|X^p]$ $n \times p$
- $(y^1, \ldots, y^q)$, q responses, $Y = [Y^1|\ldots|Y^q]$ $n \times q$

Both sample matrices are standardized

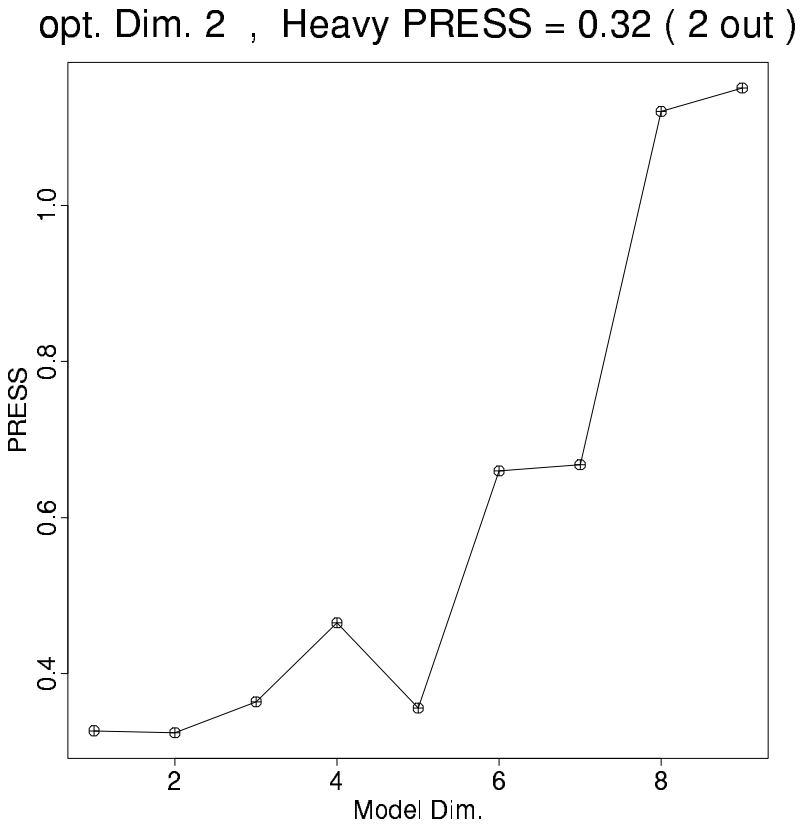
$n = 24$ orange juices: A, B, C, ..., X
$p = 10$, $q = 1$,

PREDICTORS	sensorial RESPONSE
COND (Conductivity) SiO2 Na K Ca Mg Cl HCO3 SO4 Sum	Heavy

collinearity : $SUM = SiO2 + ... + SO4$

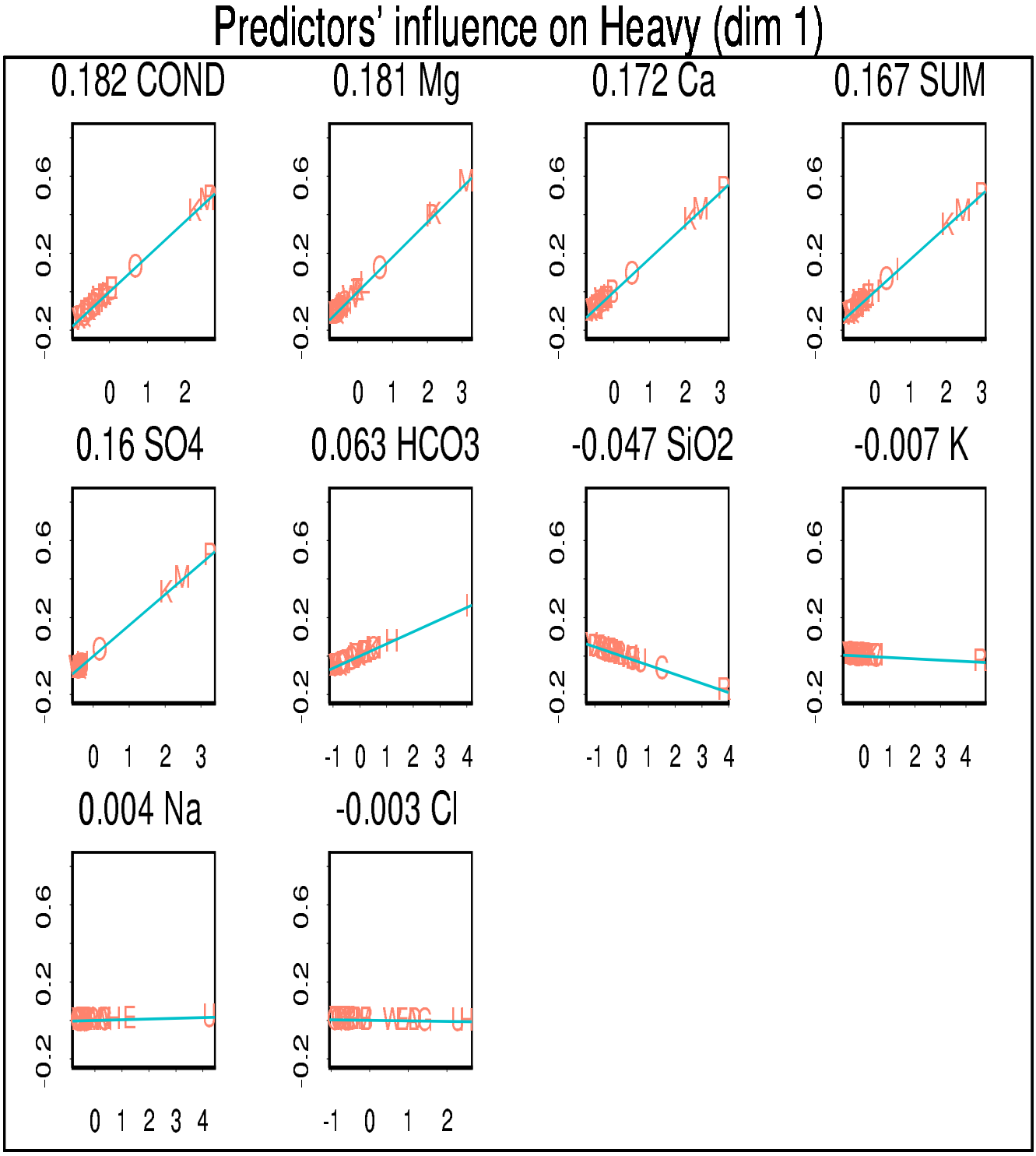

biplots between Heavy and the predictors with 'lowess'
COND
SiO2
Na
K
Ca
Mg
Cl
SO4
HCO3
SUM

Linear PLS on the orange juice data



$R^2(1) = 0.754 \qquad PRESS(1) = 0.32$

Retained dimension : $k = 1$


Predictors' influence on Heavy (dim 1)
0.182 COND
0.181 Mg
0.172 Ca
0.167 SUM
0.16 SO4
0.063 HCO3
-0.047 SiO2
-0.007 K
0.004 Na
-0.003 Cl

The main effects additive PLSS model

The centered coding matrix of X being $B = [B_1|\ldots|B_p]$

PLS through Splines (PLSS) is defined as

$$\mathbf{PLSS}(X, Y) \equiv \mathbf{PLS}(B, Y)$$

- Tuning parameters: the spline spaces for each predictor

  the model dimension $k \mapsto$ crossvalidation

- The PLSS additive model: $j = 1, \ldots, q$,

$$\widehat{y}_k^j = s_k^{j,1}(x^1) + \ldots + s_k^{j,p}(x^p)$$

The method inherits the advantages of PLS and B-splines:

♡ if $k = rank(B)$ then PLSS(X, Y) = LSS(X, Y) if it exists
♡ PLSS$(X, Y = B) \equiv$ PLS(B, B) = NL-PCA(X), [7, Gifi]
♡ efficient with low ratio observations/(column dimension of B)
♡ efficient in the multi-collinear context for predictors (concurvity)
♡ accept null columns of B (knots in empty regions)
♡ robust against extreme values of predictors (local polynomials)
♠ no automatic procedure for choosing spline parameters
♠ no interaction terms involved

Choosing the tuning parameters

Recall the tuning parameters:

1. The spline space for each predictor:
   degree, number and location of knots
2. the number of PLS components

1. Two strategies for choosing the spline spaces

- **The ascending strategy**
  - First, take $d = 1$ with no knots $(K = 0) \mapsto$ linear model.
  - Increase the degree d, keeping $K = 0$, $\mapsto$ polynomial model.
  - for fixed d, add knots, $\mapsto$ local polynomial model

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

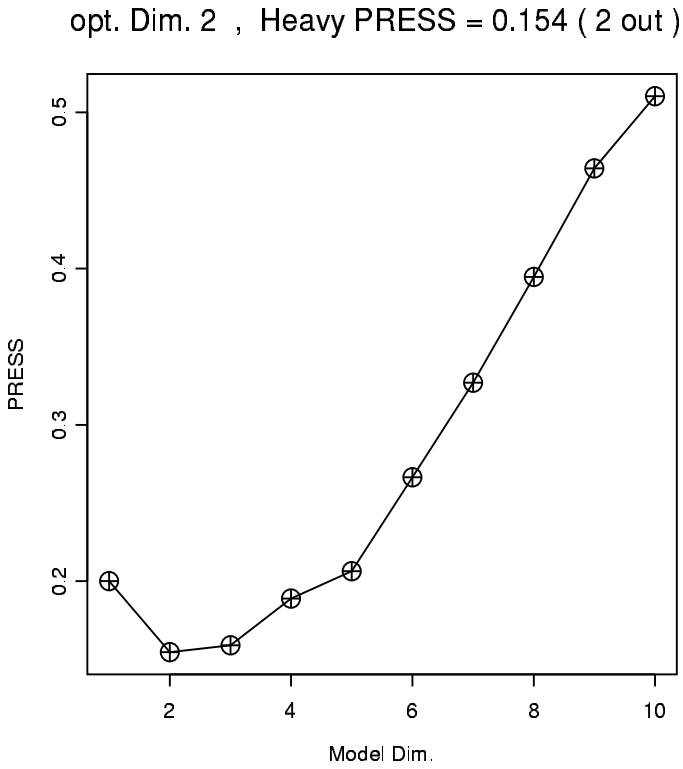
- **The descending strategy**
  - First, take a high degree, $d = 3$, and more knots than necessary.
  - Remove superfluous knots and decrease the degree as much as possible

2. To stop a strategy : find a balance between
thriftiness (k and the total spline dimension) and goodness-of fit, $R^2(k)$, and prediction, $PRESS(k)$.

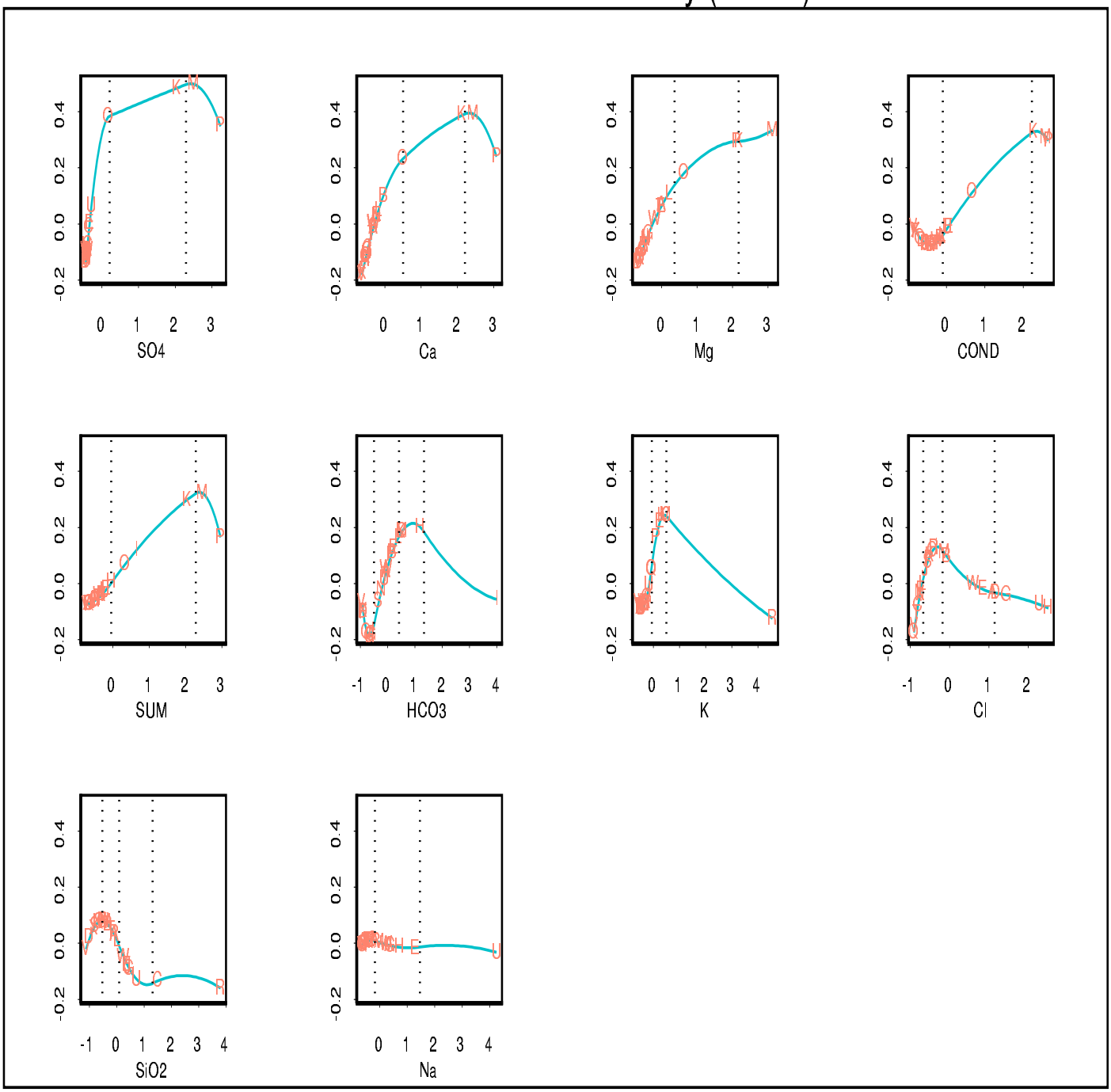
PLSS on the orange juice data

The selected knots for the predictors

COND	SiO_2	Na	K	Ca	Mg	Cl	SO_4	HCO_3	Sum
400	10	10	2.5	160	40	4	400	100	600
1600	20	40	5	400	110	11	1700	300	2600
	40					30		500	

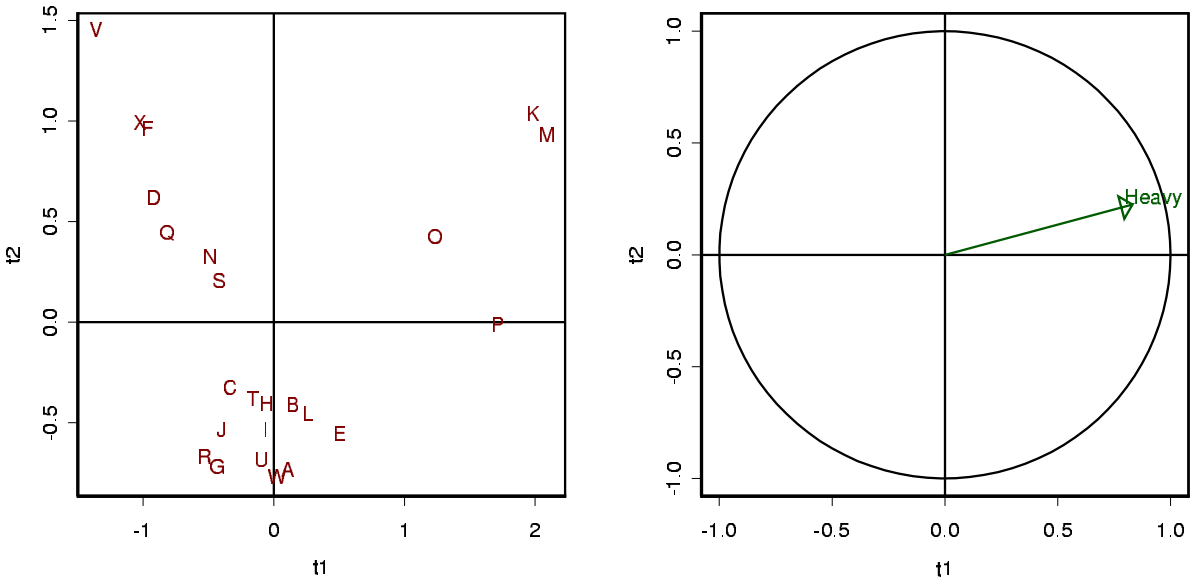


$R^2(2) = 0.917 \qquad PRESS(2) = 0.154$


Predictors' influence on Heavy (2 dim.)
SO4
Ca
Mg
COND
SUM
HCO3
K
Cl
SiO2
Na

A nonlinear look at data

On can use (t^i, t^j) and/or (t^{*i}, t^{*j}) scatter plots to look at data.

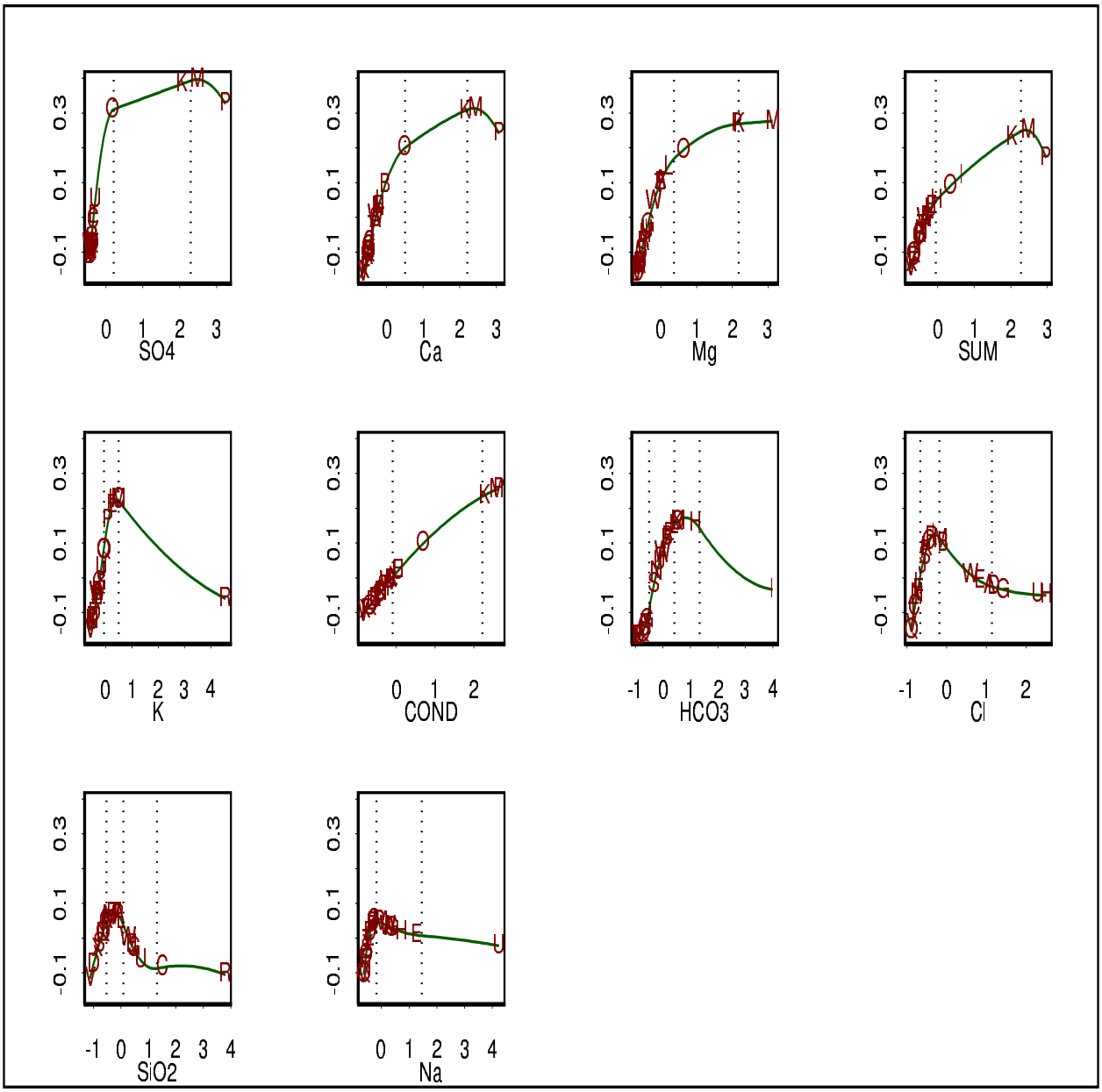


A latent variable is an additive spline function of the predictors:

$$t^i = Bw^{*i} = \sum_{j=1}^{p} B_j w_j^{*i}$$

where w_j^{*i} is the sub-vector of w^{*i} associated to the block B_j.

Notice that (t^{*i}, t^{*j}) is the representation based on projections.


Predictors' influence on t1
SO4
Ca
Mg
SUM
K
COND
HCO3
Cl
SiO2
Na

PLSS with bivariate interactions

[4, Durand & Lombardo]

The centered main effects + interactions coding matrix:

$$B = [B^1|\ldots|B^p\,\|\ldots|B^{i,i'}|\ldots]$$

where (i, i') belong to the set $\mathcal{I}$ of accepted couples of interactions

$$\mathbf{PLSS}(X, Y) \equiv \mathbf{PLS}(B, Y)$$

q simultaneous PLSS models casted in the ANOVA decomposition

$$j = 1, \ldots, q, \qquad \widehat{y}_k^j = \sum_{i=1}^{p} s_k^{j,i}(x^i) + \sum_{(i,i')\in\mathcal{I}} s_k^{j,ii'}(x^i, x^{i'})$$

PLSS with interactions shares the preceding ♡ properties

♠ no automatic procedure for choosing spline parameters

Selection of bivariate interactions

0 Preliminary phase: the main effects model.

Decide on the spline parameters as well as on k for the main effects model (m): denote $PRESS_m(k)$ and $R^2_m(k)$.

1 Individual evaluation of interaction terms.

Each interaction i is separately added to the main effects.

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

Eliminate interactions such that $CRIT(k) < 0$

Order decreasingly the accepted candidate interactions.

2 Stepwise building-model stage.

Set $PRESS_0 = PRESS_m(k)$ and $i = 0$.

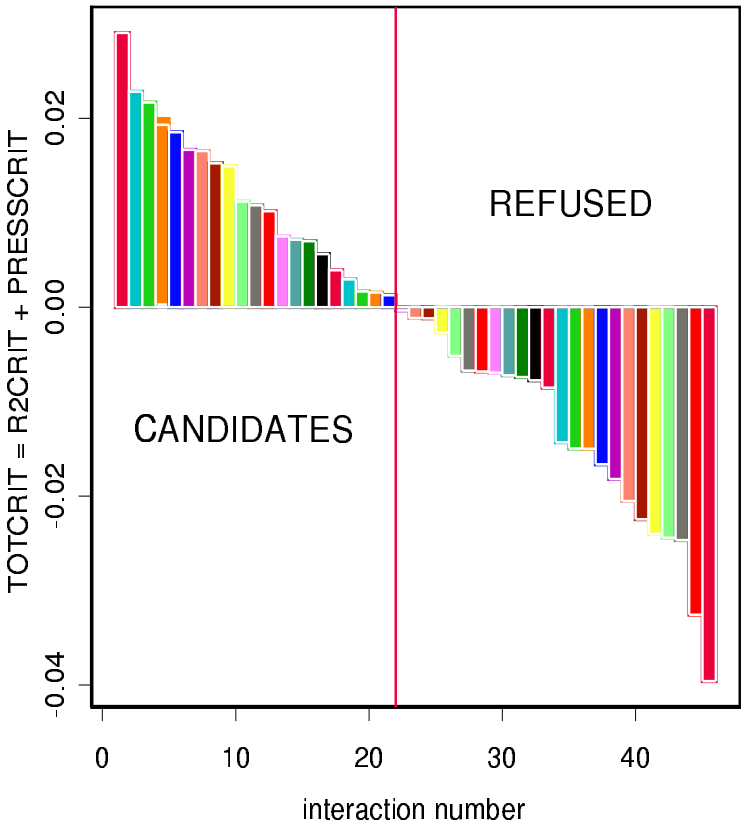
REPEAT

- $i \leftarrow i + 1$
- add interaction i and index the new model with i
- $PRESS_i \leftarrow$ optimal $PRESS$ among all dimensions

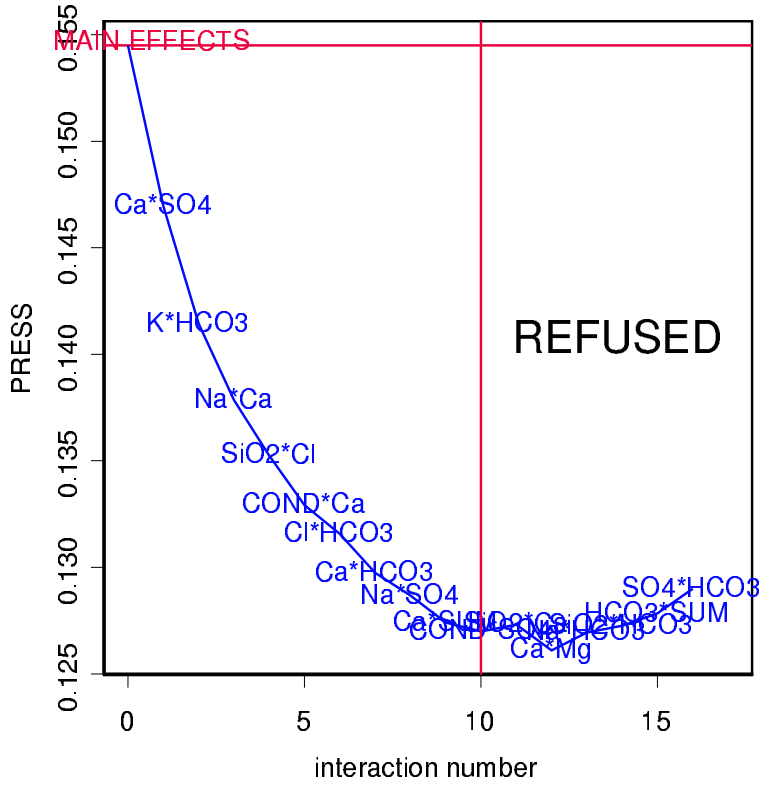
UNTIL $(PRESS_i > PRESS_{i-1})$

PLSS with interactions on the orange juice data

Phase 1: Evaluation of separate interactions (45 possible)



Phase 2: Stepwise model-building stage

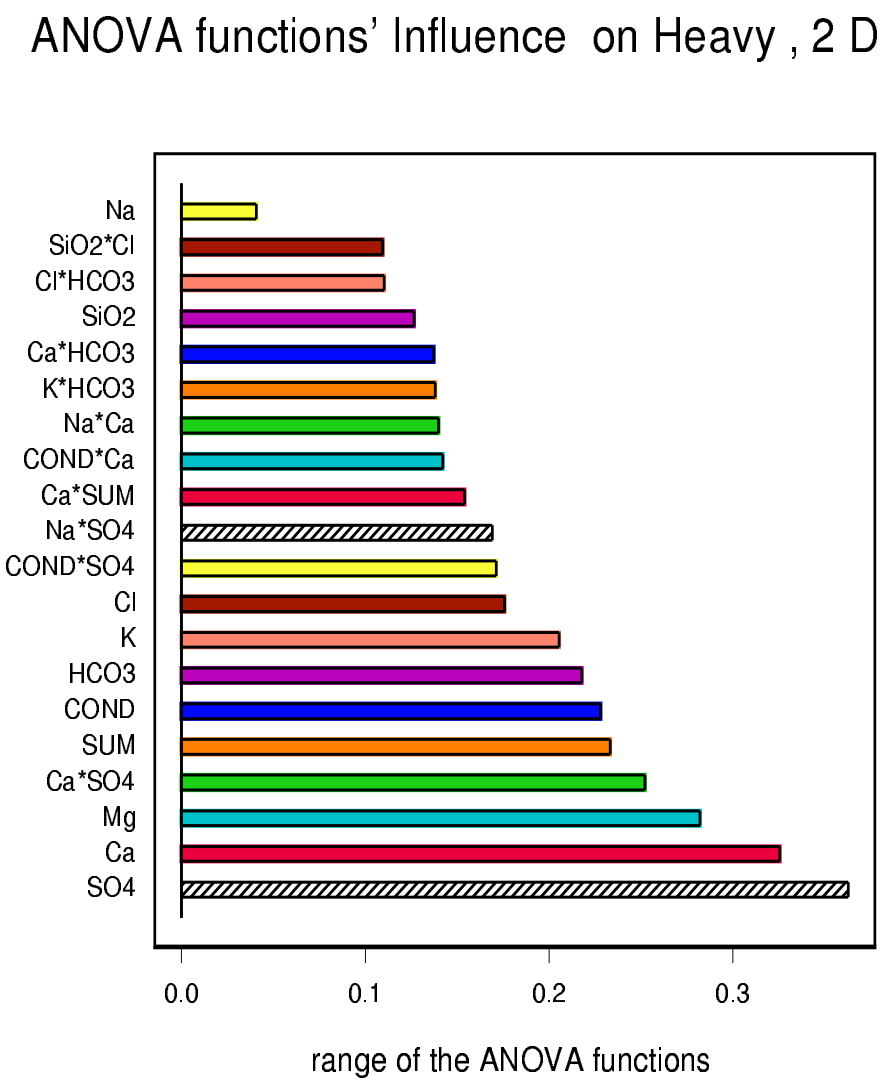


From linear PLS to PLSS with bivariate interactions:

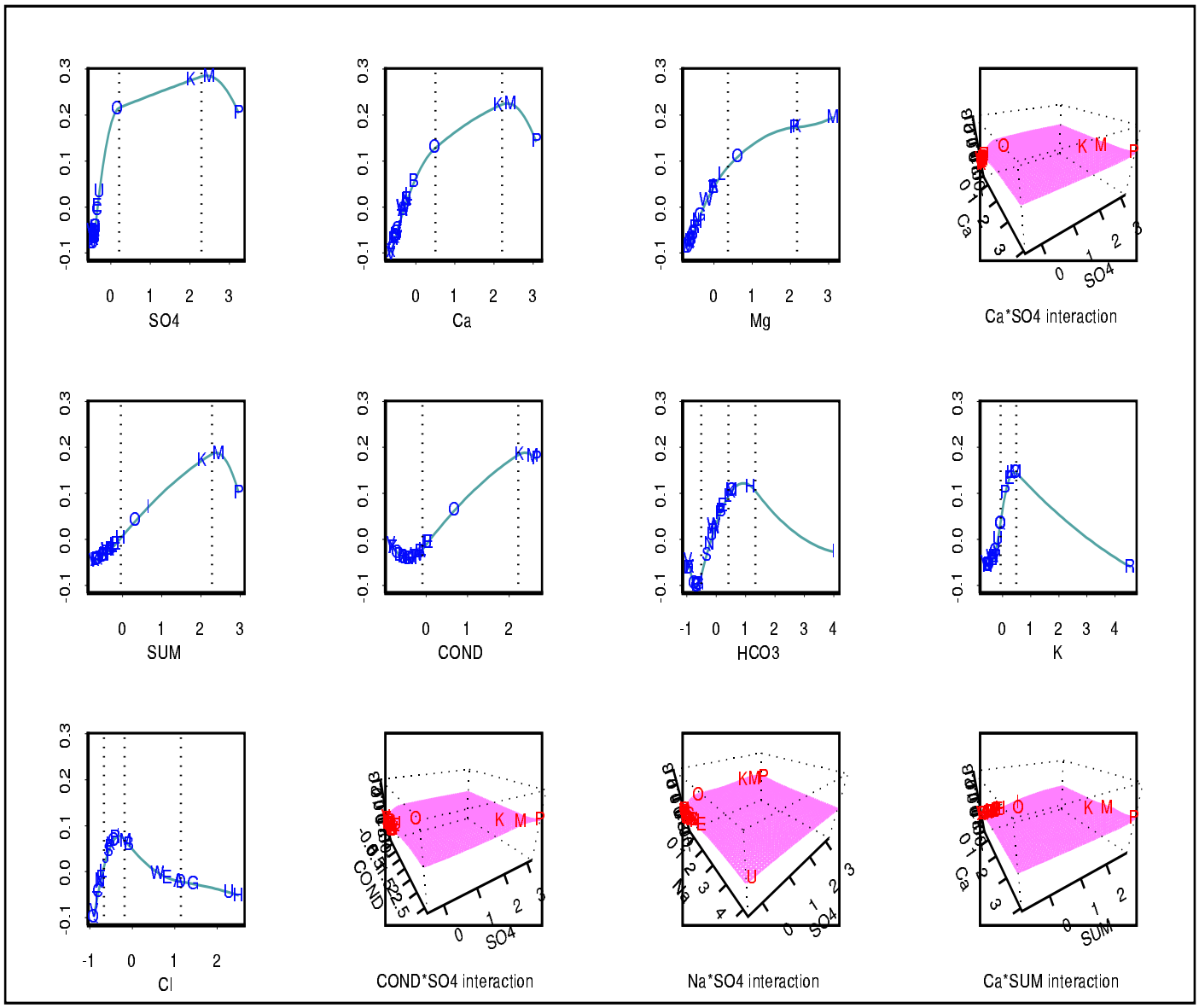
Goodness of fit and Prediction:

	PLS	$PLSS_m$	$PLSS_i$
Dim.	1	2	2
R^2	0.754	0.917	0.930
$PRESS$	0.32	0.154	0.127

Ordering and Selecting the ANOVA functions



ANOVA functions for Heavy (2 Dim.)



Section V.

Open problems with PLS and Splines

- **Revisiting D. A. methods through PLS and splines**
- **Using splines in the calibration problem**
- **Optimal knots, number and location, in the multivariate context**
- **Selection of deeper interaction levels**
- **Nonlinear multi-blocks approach for exploration and prediction**
- **Neural networks approach through splines and PLS**

Section V.

Bibliography

References

[1] C. De Boor. *A Practical Guide to Splines*, Springer-Verlag, Berlin, 1978.

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

[3] J. F. Durand. *Éléments de Calcul Matriciel et d'Analyse Factorielle de Données*, Département de Mathématiques, Université Montpellier II, 2002.

[4] J. F. Durand and R. Lombardo. *Interactions terms in nonlinear PLS via additive spline transformations*. In press in "Studies in Classification, Data Analysis, and Knowledge Organization", Springer-Verlag.

[5] European Courses in Advanced Statistics. *Methods for Multidimensional Data Analysis*, Dipartimento di Matematica e Statistica, Università di Napoli, 1987.

[6] J.H. Friedman. *Multivariate Adaptive Regression Splines, (with discussion)*. The Annals of Statistics, 19, 1-123, 1991.

[7] A. Gifi. *Non Linear Multivariate Analysis*, J. Wiley & Sons, Chichester, 1990.

[8] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley & Sons, Chichester, 1988

[9] G. Mazerolles, G. Duboz and S. Hugot. *Détermination des taux d'humidité et de matière grasse de fromages de type pâte presseé par spectroscopie proche infrarouge en mode transmission*, Lait 80, 371-379, © INRA, EDP Sciences, 2000.

[10] N. Molinari, J.F. Durand and R. Sabatier. *Bounded Optimal knots for Regression Splines*. Computational Statistics & Data Analysis. In press, 2003.

[11] C.J. Stone. *Additive regression and other nonparametric models*. The Annals of Statistics, 13, 689-705, 1985.

[12] M. Tenenhaus. *La régression PLS, Théorie et Applications*, Technip, Paris, 1998.

[13] H. Wold. *Estimation of principal components and related models by iterative least squares*. In Multivariate Analysis, (Eds.) P.R. Krishnaiah, New York: Academic Press, 391-420, 1966.

[14] S. Wold., H. Martens and H. Wold. *The mutivariate calibration problem in chemistry solved by PLS method*. In: A. Ruhe, B. Kagstrom (Eds), Lecture Notes in Mathematics, Proceedings of the Conference on Matrix Pencils, Springer-Verlag, Heidelberg, 286-293, 1983.