Les Méthodes PLS

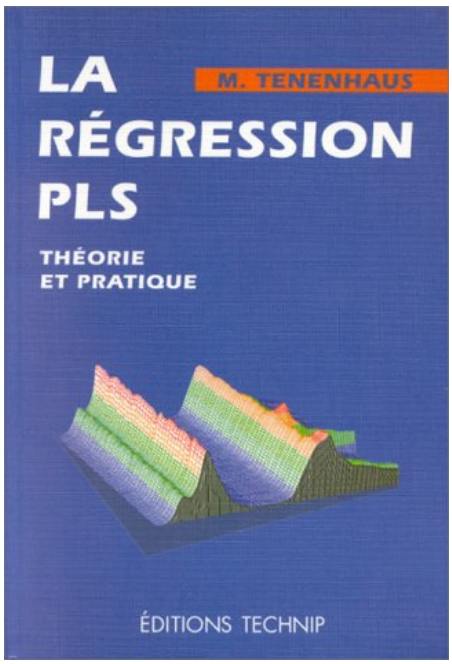
Michel Tenenhaus
tenenhaus@hec.fr

CHAMBRE DE COMMERCE ET D'INDUSTRIE DE PARIS

Les méthodes PLS
initiées par Herman et Svante Wold

I. NIPALS (Nonlinear Iterative Partial Least Squares)
II. Régression PLS (Partial Least Squares Regression)
III. Analyse discriminante PLS
IV. SIMCA (Soft Independent Modelling by Class Analogy)
V. PLS Path Modelling (Modélisation de relations structurelles sur variables latentes)
VI. Régression logistique PLS
VII. Régression linéaire généralisée PLS

2/200



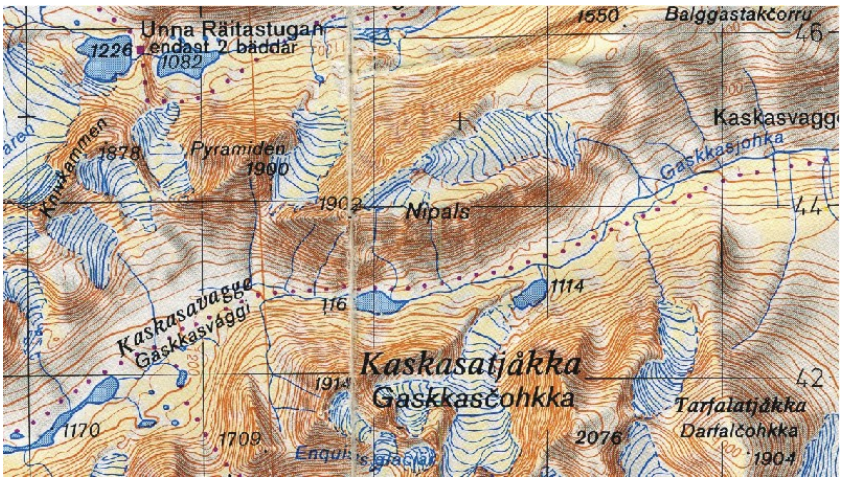
3/200

Les méthodes PLS

I. NIPALS (Nonlinear Iterative Partial Least Squares)

4/200

Le mont NIPALS



5/200

Le mont NIPALS

6/200

La méthode NIPALS

Analyse en composantes principales

- Possibilité de données manquantes.
- Validation croisée pour choisir le nombre de composantes.
- Identification des outliers avec
  - une carte de contrôle des observations,
  - des tests sur les écarts au modèle de l'ACP.

7/200

Utilisation de NIPALS : Exemple voitures

Modèle	Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur
Honda Civic	.	90	174	850	369	166
Renault 19	1721	.	180	965	415	169
Fiat Tipo	1580	83	.	970	395	170
⋮						
Citroën AX Sport	1294	95	184	730	350	.

Il y a une observation manquante par modèle !!!

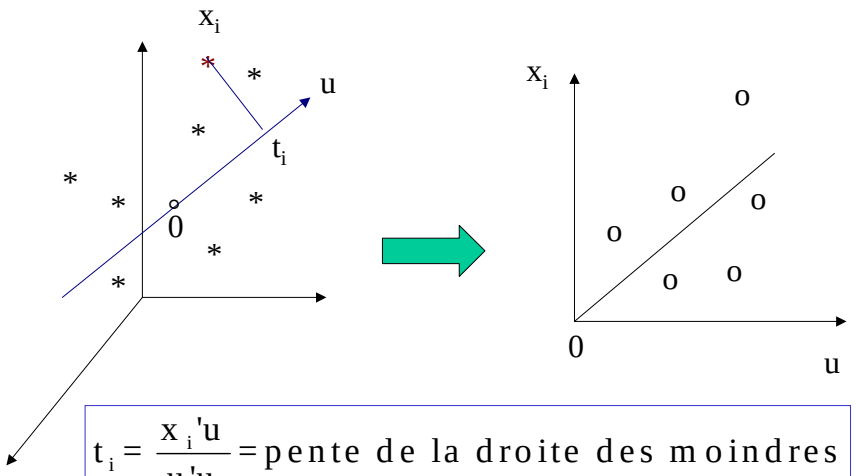
8/200

Le principe de NIPALS

Comment projeter un point avec données manquantes ?

9/200

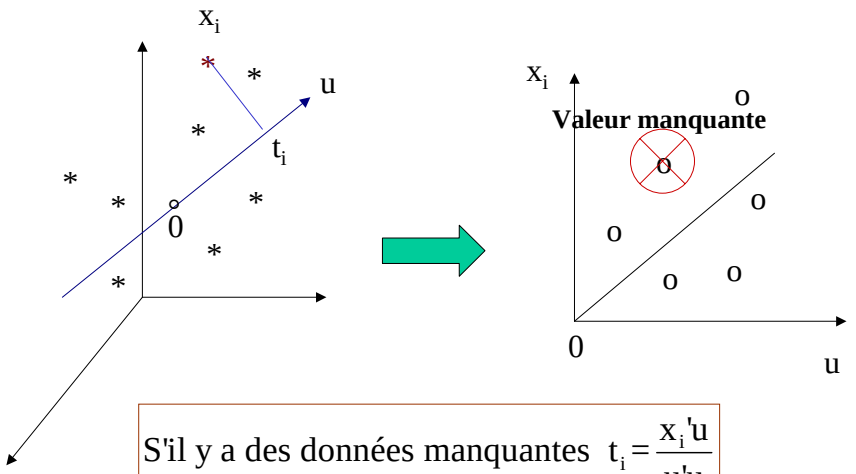
Projection sur un axe



$$t_i = \frac{x_i'u}{u'u} = \text{pente de la droite des moindres carrés sans constante de } x_i \text{ sur } u.$$

10/200

Projection d'un point avec données manquantes sur un axe



S'il y a des données manquantes $t_i = \frac{x_i'u}{u'u}$

est calculé sur les données disponibles.

11/200

L 'algorithme NIPALS

Recherche des composantes principales

Données :

$X = \{x_{ij}\}$ tableau $n \times k$,

$\mathbf{x}_j$ = variable j

x_i = observation i

Modèle de l 'ACP :

$X = t_1 p_1' + \ldots + t_k p_k'$

avec (1) $p_1, \ldots, p_k$ orthonormés

et (2) $t_1, \ldots, t_k$ orthogonaux

12/200

L 'algorithme NIPALS
Recherche de la première composante principale

- Modèle : $X = t_1 p_1' +$ résidu, avec p_1 normé
- Algorithme : les équations de base

  (1) Si t_1 connu, calcul de p_{1j} par régression simple :

  $$\mathbf{x}_j = p_{1j} t_1 + \text{résidu}$$

  (2) Normalisation de $p_1 = (p_{11},\ldots,p_{1k})$

  (3) Si p_1 connu, calcul de t_{1i} par régression simple :

  $$x_i = t_{1i} p_1 + \text{résidu}$$
- Algorithme : fonctionnement
  - Prendre $t_1 = \mathbf{x}_1$, puis itérer sur (1), (2), (3).
  - Si données manquantes, faire les calculs sur toutes les données disponibles.

13/200

L 'algorithme NIPALS
Recherche des autres composantes principales

- La première étape donne :

$$X = t_1 p_1' + X_1$$

- On répète les opérations précédentes sur la matrice des résidus X_1 de la régression de X sur t_1.
- On obtient : $X_1 = t_2 p_2' + X_2$

$$\text{et} \quad X = t_1 p_1' + t_2 p_2' + X_2$$

- On obtient de même les autres composantes.

14/200

$RESS_h$ et $PRESS_h$

A chaque étape on étudie la reconstitution du tableau X :

$$\hat{X} = t_1 p_1' + t_2 p_2' + \ldots + t_h p_h'$$

Residual Sum of Squares : $RESS_h = \sum_{i,j} (x_{ij} - \hat{x}_{ij})^2$

Les cases de X sont partagées en G groupes, et on réalise G factorisations en enlevant à chaque fois un seul des groupes.

Voir opti...CV...

Predicted Residual Sum of Squares :

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

Ress = pouvoir explicatif

Press = pouvoir predictif

où $\hat{x}_{(-ij)}$ est calculé dans l'analyse réalisée sans le groupe contenant la case (i,j).

15/200

L 'algorithme NIPALS
Choix du nombre de composantes

- On choisit le nombre de composantes principales par validation croisée.
- La composante t_h est retenue selon la règle R1 si

$$Q^2 = 1 - \frac{PRESS_h}{RESS_{h-1}} \geq \text{limite}$$

Cette règle conduit à des composantes globalement significatives.

16/200

Q^2(cum) et R^2(validation croisée)

$$[Q^2_{cum}]_h = 1 - \prod_{a=1}^{h} \frac{PRESS_a}{RESS_{a-1}}$$

peu différent de

$$R^2_{\text{validation croisée}} = 1 - \frac{\text{PRESS}_h}{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2}$$

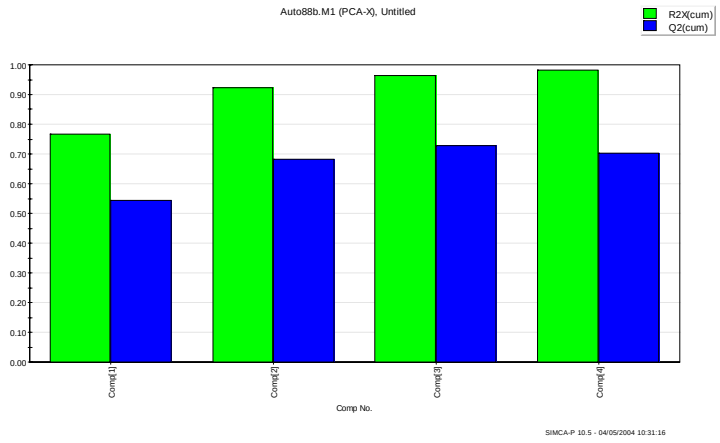
La composante h est retenue si :

$[Q^2_{cum}]_h$ est nettement supérieur à $[Q^2_{cum}]_{h-1}$

CONSEIL : Modèle à h composantes acceptable si $[Q^2_{cum}]_h > 0.5$

17/200

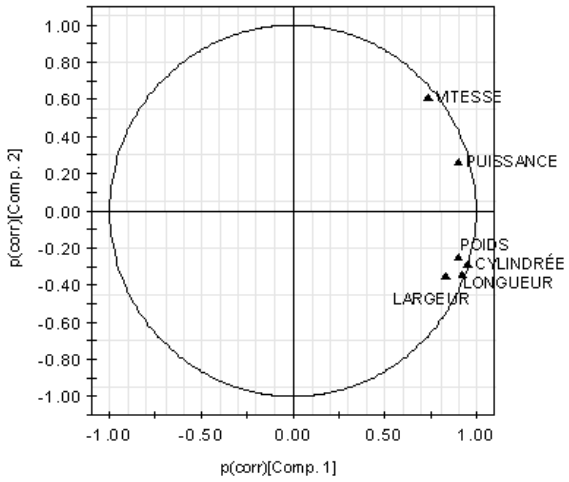
Utilisation de NIPALS : Exemple voitures



La validation croisée conduit à deux composantes globalement significatives (critère R1).

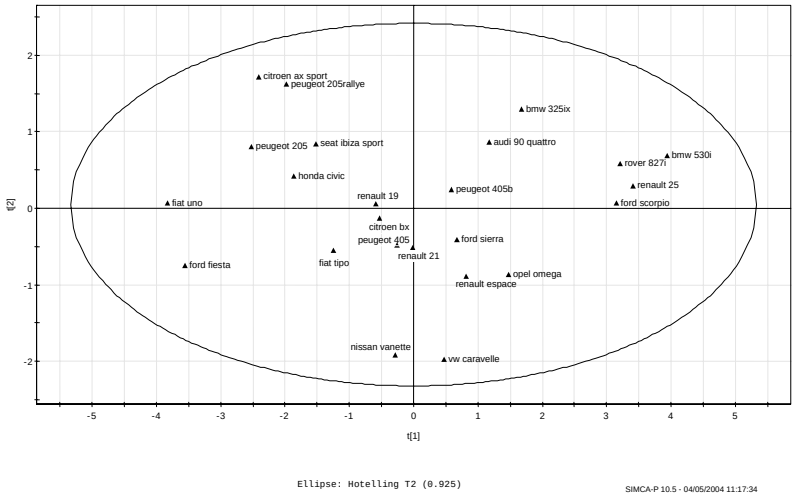
18/200

NIPALS : Exemple Voitures
Cercle des corrélations



19/200

NIPALS : Exemple Voitures
Carte des voitures (les 2 premières "composantes principales")

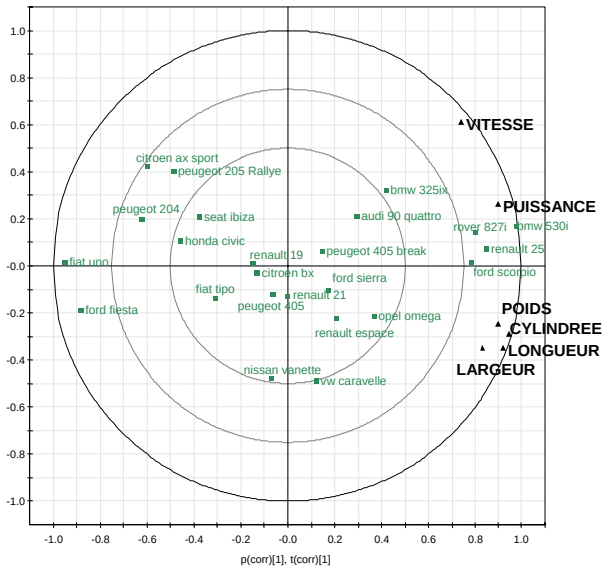


Ellipse: Hotelling T2 (0.925)

SIMCA-P 10.5 - 04/05/2004 11:17:34

20/200

Loadings Bi-plot



21/200

Statistiques, vecteurs propres et composantes principales

	N	% MisVal	Mean	Std.dev
CYLINDREE	20	16.6667	1887.70	459.31
PUISSANCE	20	16.6667	112.05	36.14
VITESSE	20	16.6667	181.95	24.84
POIDS	20	16.6667	1112.75	238.49
LONGUEUR	20	16.6667	421.70	42.78
LARGEUR	20	16.6667	168.40	7.34

	p[1]	p[2]
CYLINDREE	0.49	-0.07
PUISSANCE	0.44	0.35
VITESSE	0.37	0.62
POIDS	0.39	-0.34
LONGUEUR	0.39	-0.35
LARGEUR	0.36	-0.50

	T1	T2
honda civic	-1.86	0.43
renault 19	-0.59	0.06
fiat tipo	-1.24	-0.55
peugeot 405	-0.26	-0.48
renault 21	-0.02	-0.51
citroen bx	-0.53	-0.13
bmw 530i	3.94	0.69
rover 827i	3.21	0.59
renault 25	3.42	0.30
opel omega	1.48	-0.86
peugeot 405b	0.58	0.25
ford sierra	0.68	-0.41
bmw 325ix	1.68	1.29
audi 90 quattro	1.17	0.86
ford scorpio	3.15	0.07
renault espace	0.82	-0.89
nissan vanette	-0.29	-1.92
vw caravelle	0.48	-1.97
ford fiesta	-3.56	-0.75
fiat uno	-3.83	0.07
peugeot 205	-2.52	0.81
peugeot 205 rallye	-1.97	1.62
seat ibiza sxi	-1.52	0.85
citroen ax sport	-2.41	1.72

22/200

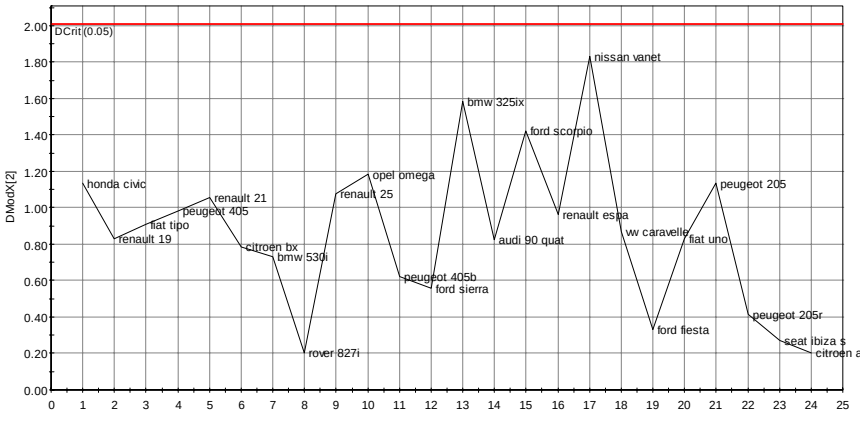
Reconstitution des données avec deux composantes

Obs ID (Primary)	CYLINDREE	PUISSANCE	VITESSE	POIDS	LONGUEUR	LARGEUR
honda civic	1460	88	172	905	384	162
renault 19	1754	103	177	1053	411	167
fiat tipo	1627	85	162	1041	409	167
peugeot 405	1844	102	172	1127	424	169
renault 21	1900	105	174	1153	429	170
citroen bx	1774	102	175	1074	415	167
bmw 530i	2744	183	228	1424	478	176
rover 827i	2585	170	220	1365	467	175
renault 25	2640	170	218	1407	475	176
opel omega	2245	125	182	1321	459	175
peugeot 405b	2010	124	191	1147	428	169
ford sierra	2052	118	182	1210	439	172
bmw 325ix	2221	155	217	1164	431	168
audi 90 quattro	2122	141	206	1152	429	168
ford scorpio	2588	163	212	1401	474	176
renault espace	2098	114	176	1262	449	174
nissan vanette	1885	83	150	1242	445	175
vw caravelle	2057	95	156	1318	459	177
ford fiesta	1117	46	138	842	373	162
fiat uno	1031	52	148	750	356	158
peugeot 205	1300	82	172	812	367	159
peugeot 205 rallye	1396	101	189	797	364	157
seat ibiza sxi	1522	99	181	902	384	161
citroen ax sport	1296	96	187	748	356	156

23/200

NIPALS : Identification des outliers

Carte de contrôle des distances au modèle normalisées

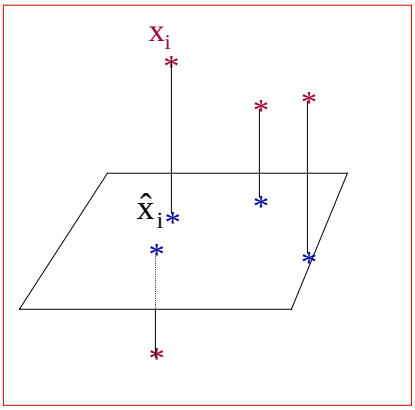


Dcrit [2] = 2.00746, Normalized distances, Non weighted residuals

Simca-P 8.0 by Umetrics AB 2000-05-30 19:00

24/200

Calcul de la limite de contrôle



Propriété :

$$DModX = \sqrt{\frac{d^2(x_i,\hat{x}_i)}{\frac{1}{n}\sum_{i=1}^{n} d^2(x_i,\hat{x}_i)}} \approx \sqrt{F(k_1,k_2)}$$

calculé si nb de données > nb de CP

Limite de contrôle au risque α :

$$\sqrt{F_{1-\alpha}(k_1,k_2)}$$

25/200

Probabilité d'appartenir au modèle

Test : H_0 : l'observation i appartient au modèle de l'ACP
H_1 : l'observation i n'appartient pas au modèle

Décision : On rejette H_0 au risque α de se tromper si

$$DModX \geq \sqrt{F_{1-\alpha}(k_1,k_2)}$$

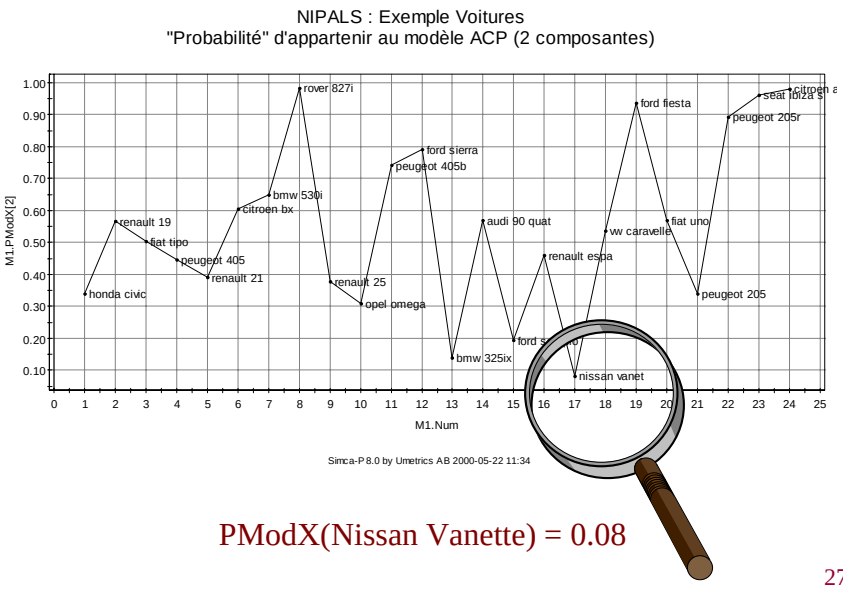
Niveau de signification ou « probabilité d'appartenir au modèle » :

Plus petit α conduisant au rejet de H_0

= Prob ($F(k_1,k_2) \geq DModX^2$)

L'individu i est exactement sur la limite de contrôle $DCrit(\alpha_{min})$

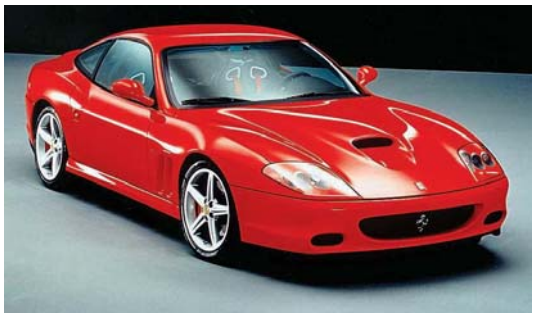
26/200



PModX(Nissan Vanette) = 0.08

27/200

Ajouter la Ferrari au fichier des données



Caractéristiques de la Ferrari

Cylindrée : 4943
Puissance : 428
Vitesse : 310
Poids : 1517
Longueur : 449

28/200

Les méthodes PLS

II. Régression PLS (Partial Least Squares Regression)

29/200

La régression PLS

- Relier un bloc de variables à expliquer Y à un bloc de variables explicatives X.
- Possibilité de données manquantes.
- Il peut y avoir beaucoup plus de variables X que d'observations.
- Il peut y avoir beaucoup plus de variables Y que d'observations.
- Meilleure réponse au problème de la multicolinéarité.

30/200

La régression PLS : vocabulaire

- Régression PLS1 : un seul Y
- Régression PLS2 : plusieurs Y
  A conseiller si les Y sont corrélés entre eux
- Analyse discriminante PLS :
  - Y qualitatif transformé en variables indicatrices des modalités
  - A conseiller si Y binaire, sinon on peut peut- être faire mieux (Barker & Rayens, 2003)

31/200

Les méthodes PLS

II.1 Régression PLS1

32/200

La régression PLS1 : une idée de l'algorithme

Etape 1 : Recherche de m composantes orthogonales $t_h = Xa_h$ bien explicatives de leur propre groupe et bien corrélées à y.

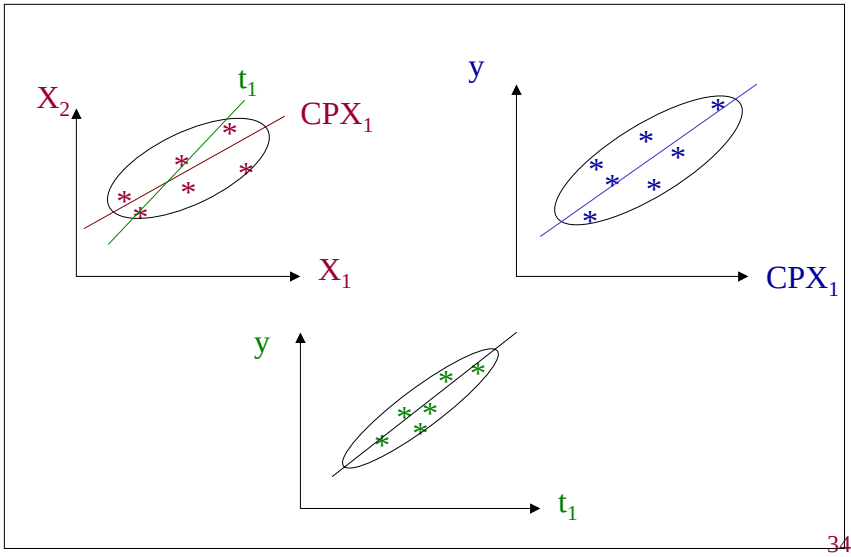
Le nombre m est obtenu par validation croisée.

Etape 2 : Régression de Y sur les composantes PLS t_h.

Etape 3 : Expression de la régression en fonction de X.

33/200

Objectif de l'étape 1 de la régression PLS1



34/200

La régression PLS1 : une idée de l'étape 1 lorsqu'il n'y a pas de données manquantes

Pour chaque h = 1 à m, on recherche des composantes $t_h = Xa_h$ maximisant le critère

$$\text{Cov}(Xa_h, y)$$

sous des contraintes de norme ($\|a_h\| = 1$) et d'orthogonalité entre t_h et les composantes précédentes $t_1, \ldots, t_{h-1}$.

35/200

Propriétés de la régression PLS1

De $\text{Cov}^2(Xa_h, y)$

$= \text{Cor}^2(Xa_h, y) * \text{Var}(Xa_h) * \text{Var}(y)$

on déduit que la régression PLS1 réalise un compromis entre la régression multiple de y sur X et l'analyse en composantes principales de X.

36/200

Régression PLS1: Étape 1

1. Calcul de la première composante PLS t_1 :

$$t_1 = Xa_1 = \sum_j \underbrace{cor(y, x_j)}_{\text{cor avec y > 0}} \times x_j$$

Les x_j sont centrés-réduits. Sinon, remplacer la corrélation par la covariance.

2. Normalisation du vecteur $a_1 = (a_{11},\ldots,a_{1k})$
3. Régression de y sur $t_1 = Xa_1$ exprimée en fonction des x
4. Calcul des résidus y_1 et X_1 des régressions de y et X sur t_1 :
   - $y = c_1 t_1 + y_1$
   - $X = t_1 p_1' + X_1$

Régression PLS1: Étape 2

1. Calcul de la deuxième composante PLS t_2 :

$$t_2 = X_1 b_2 = \sum_j \text{cov}(y_1, x_{1j}) \times x_{1j}$$

2. Normalisation du vecteur $b_2 = (b_{21},\ldots,b_{2k})$
3. Calcul de a_2 tel que : $t_2 = X_1 b_2 = Xa_2$
4. Régression de y_1 sur $t_2 = Xa_2$ exprimée en fonction des x
5. Calcul des résidus y_2 et X_2 des régressions de y et X_1 sur t_2 :
   - $y_1 = c_2 t_2 + y_2$
   - $X_1 = t_2 p_2' + X_2$

Régression PLS1: Étapes suivantes

- On procède de la même manière pour les autres composantes.
- D'où le modèle de régression PLS à m composantes :

$$
\begin{aligned}
y &= c_1t_1 + c_2t_2 + \ldots + c_mt_m + \text{Résidu} \\
&= c_1Xa_1 + c_2Xa_2 + \ldots + c_mXa_m + \text{Résidu} \\
&= X(c_1a_1 + c_2a_2 + \ldots + c_ma_m) + \text{Résidu} \\
&= \underbrace{b_1x_1 + b_2x_2 + \ldots + b_kx_k}_{\hat{y}} + \text{Résidu}
\end{aligned}
$$

Calcul de $RESS_h$ et $PRESS_h$ à l'étape h

Residual Sum of Squares : $RESS_h = \sum_i (y_{(h-1),i} - \hat{y}_{(h-1),i})^2$

où $\hat{y}_{(h-1),i} = c_h t_{hi}$ est la prévision de $y_{(h-1),i}$

Les observations sont partagées en G groupes, et on réalise G fois l'étape courante de l'algorithme sur y_{h-1} et X_{h-1} en enlevant à chaque fois un groupe.

Predicted Residual Sum of Squares :

Voir option CV-groups

$$PRESS_h = \sum_i (y_{(h-1),i} - \hat{y}_{(h-1),-i})^2$$

où $\hat{y}_{(h-1),-i}$ est calculé dans l'analyse réalisée sans le groupe contenant l'observation (i).

Choix du nombre de composantes

- On choisit le nombre de composantes par validation croisée.
- La composante h est retenue si

$\Rightarrow \quad [PRESS_h] \leq \gamma\,[RESS_{h-1}]$

Soit :

$$Q^2 = 1 - \frac{PRESS_h}{RESS_{h-1}} \geq 1 - \gamma$$

avec $RESS_0 = \sum (y_i - \bar{y})^2$, $1 - \gamma = 0.05$ si $n < 100$ et $= 0$ si $n \geq 100$.

41/200

Q^2(cum) et R^2(validation croisée)

$$[Q^2_{cum}]_h = 1 - \prod_{a=1}^{h} \frac{PRESS_a}{RESS_{a-1}}$$

peu différent de

$$R^2_{\text{validation croisée}} = 1 - \frac{PRESS_h}{\sum_i (y_i - \bar{y})^2}$$

La composante h est retenue si :

$[Q^2_{cum}]_h$ est nettement supérieur à $[Q^2_{cum}]_{h-1}$

Modèle à h composantes acceptable si $[Q^2_{cum}]_h > 0.5$

42/200

Variable Importance in the Prediction (VIP)

- Composantes PLS : $t_h = X_{h-1} b_h$, avec $\|b_h\| = 1$
- Importance de la variable x_j (j=1,…, p) pour la prédiction de y dans un modèle à m composantes :

$$VIP_{mj} = \sqrt{\frac{p}{\sum_{h=1}^{m} cor^2(y,t_h)} \sum_{h=1}^{m} cor^2(y,t_h) b_{hj}^2}$$

- Moyenne des carrés des VIP sur les variables = 1
- Variable importante pour la prédiction si VIP > 0.8

Régression PLS1 : Exemple Voitures

Problèmes : multicolinéarité, données manquantes

Données complètes

Modèle	Prix	Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur
Honda Civic	83700	1396	90	174	850	369	166
Renault 19	83800	1721	92	180	965	415	169
Fiat Tipo	70100	1580	83	170	970	395	170
⋮							
Citroën AX Sport	66800	1294	95	184	730	350	160

Données incomplètes

Modèle	Prix	Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur
Honda Civic	83700	.	90	174	850	369	166
Renault 19	83800	1721	.	180	965	415	169
Fiat Tipo	70100	1580	83	.	970	395	170
⋮							
Citroën AX Sport	66800	1294	95	184	730	350	.

Régression multiple sur les données complètes

$R^2 = 0.847$, $F = 15.730$ Sig. = 0.0001

Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
		B	Std. Error			Lower Bound	Upper Bound
1	(Constant)	12070.406	194786.6	.062	.951	-398893.309	423034.120
	CYLINDRE	-1.936	33.616	-.058	.955	-72.860	68.988
	PUISSANC	1315.906	613.510	2.145	.047	21.512	2610.299
	VITESSE	-472.507	740.319	-.638	.532	-2034.443	1089.428
	POIDS	45.923	100.047	.459	.652	-165.158	257.005
	LONGUEUR	209.653	504.152	.416	.683	-854.014	1273.319
	LARGEUR	-505.429	1501.589	-.337	.741	-3673.505	2662.648

a. Dependent Variable: PRIX

45/200

Corrélations entre les variables

Correlation Matrix

	Correlation						
	PRIX	CYLINDRE	PUISSANC	VITESSE	POIDS	LONGUEUR	LARGEUR
PRIX	1.000	.852	.891	.720	.813	.747	.611
CYLINDRE	.852	1.000	.861	.693	.905	.864	.709
PUISSANC	.891	.861	1.000	.894	.746	.689	.552
VITESSE	.720	.693	.894	1.000	.491	.532	.363
POIDS	.813	.905	.746	.491	1.000	.917	.791
LONGUEUR	.747	.864	.689	.532	.917	1.000	.864
LARGEUR	.611	.709	.552	.363	.791	.864	1.000

Tolerance
= $1-R^2$(X,autres X)

VIF = 1/Tolerance)

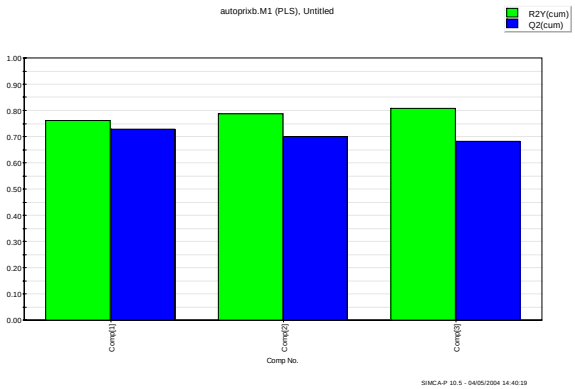
Problème si VIF > 3,
inacceptable si VIF > 10

Coefficients

	Collinearity Statistics	
	Tolerance	VIF
CYLINDRE	.094	10.608
PUISSANC	.052	19.071
VITESSE	.085	11.738
POIDS	.056	17.880
LONGUEUR	.068	14.631
LARGEUR	.225	4.449

46/200

Régression PLS sur les données incomplètes
Choix du nombre de composantes



On retient une composante PLS

Régression PLS sur les données **incomplètes**

$R^2 = 0.761$

Équation sur les données centrées-réduites (CoeffCS)

$$\frac{Prix}{\sigma(Prix)} = 2.18 + 0.183\text{Cylindrée}^* + 0.206\text{Puissance}^* + 0.146\text{Vitesse}^* + 0.165\text{Poids}^* + 0.153\text{Longueur}^* + 0.129\text{Largeur}^*$$

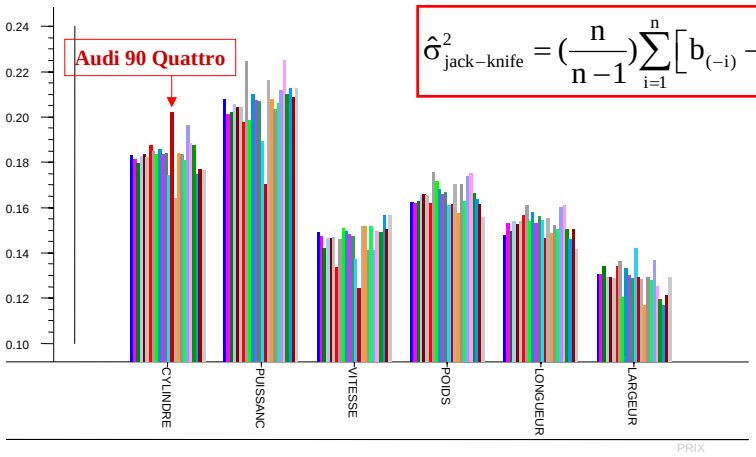
Équation sur les données d'origine (Coeff)

Prix = -316 462 + 23Cylindrée + 328Puissance + 339Vitesse + 40Poids + 205Longueur + 1007Largeur

Équation sur les données d'origine pour Y et centrées pour X (CoeffC)

Prix = 125513 + 23(Cylindrée - 1888) + 328(Puissance - 112) + 339(Vitesse - 182) + 40(Poids - 1113) + 205(Longueur - 422) + 1007(Largeur - 168)

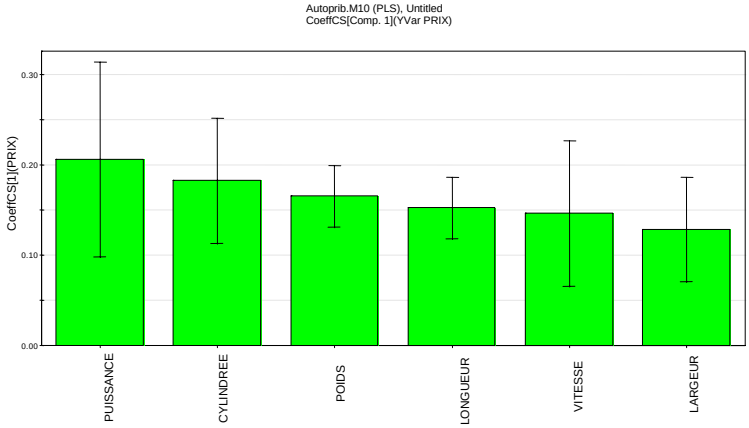
Résultats de la validation croisée sur les coefficients de régression PLS



$$\hat{\sigma}^2_{\text{jack-knife}} = \left(\frac{n}{n-1}\right)\sum_{i=1}^{n}\left[b_{(-i)} - b_{\text{total}}\right]^2$$

49/200

Intervalles de confiance jack-knife des coefficients de régression PLS



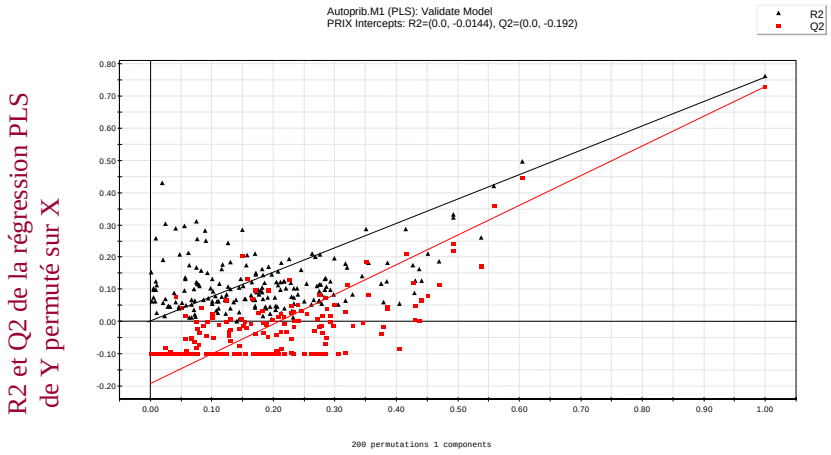
50/200

Résultats de la validation croisée sur les coefficients de régression PLS

	B	SE	Student T	p-value
Cylindrée	0.1827	0.0371	4.925	0.0001
Puissance	0.2060	0.0570	3.614	0.0005
Vitesse	0.1465	0.0430	3.407	0.0002
Poids	0.1653	0.0181	9.133	0.0001
Longueur	0.1525	0.0175	8.714	0.0001
Largeur	0.1286	0.0299	4.301	0.0001

51/200

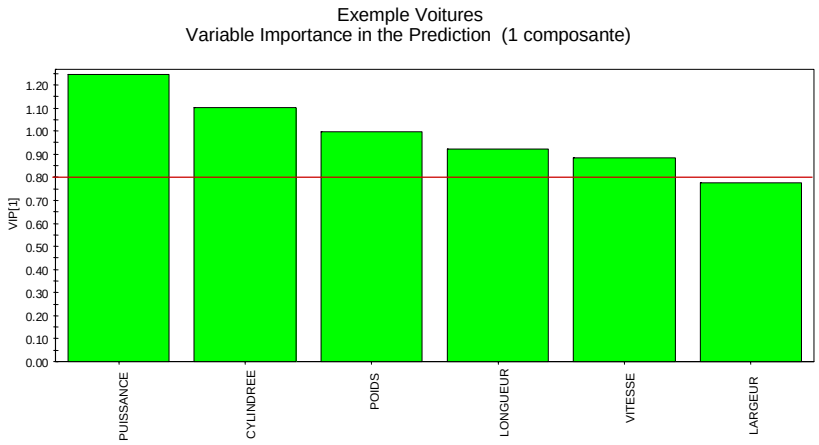
Validation globale



Corrélation entre Y observé et Y permuté

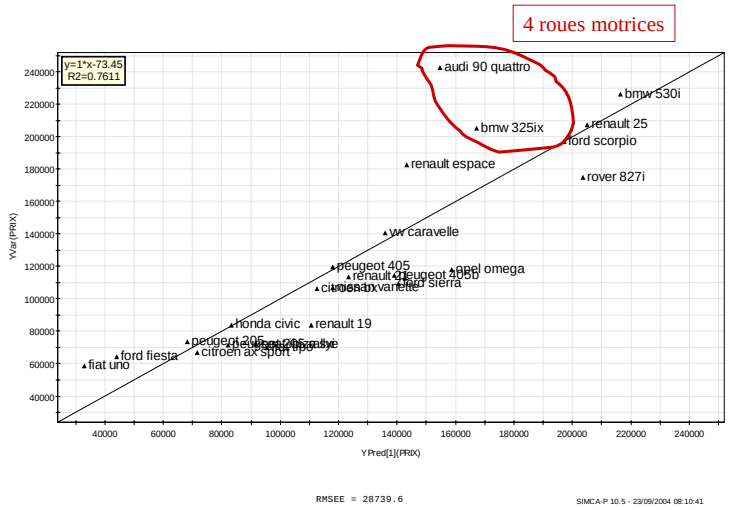
Les droites noire et rouge sont les droites des moindres carrés passant par les R2 et Q2 observés. Le modèle est validé si les ordonnées à l'origine sont proches de 0 pour R^2 et <u>surtout</u> négative pour Q^2.

52/200



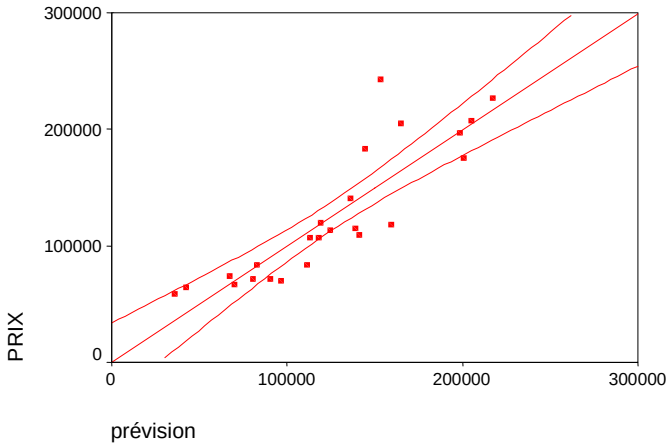
53/200

Régression PLS sur les données incomplètes



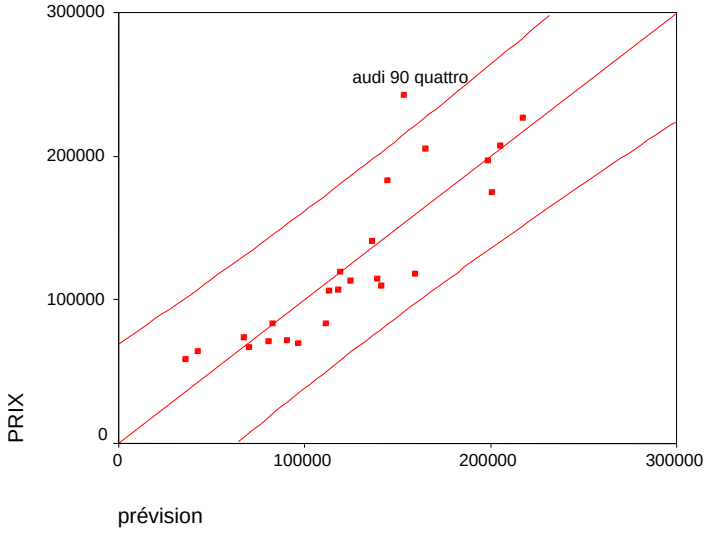
54/200

Intervalle de confiance à 95% du prix moyen
(fourni par SIMCA)



55/200

Intervalle de prévision à 95% du prix
(à calculer)



56/200

Prédiction du prix de la HONDA CIVIC (WS) (Problème : certains X sont manquants)

Prix de vente : 83 700 FF

	Caractéristiques de la Honda Civic	Caractéristiques centrées-réduites
Cylindrée	?	?
Puissance	90	-.61009
Vitesse	174	-.32011
Poids	850	-1.10172
Longueur	369	-1.23196
Largeur	166	-.32679

Prédiction du Prix de la HONDA CIVIC (WS)

Régression du Prix sur t_1 :

$$\frac{\text{Prix} - 125\,512}{57\,503} \approx \underbrace{0.4045789}_{c} \times t_1$$

Calcul de t_1 pour la HONDA CIVIC :

$$t_1(\text{Honda Civic}) = \frac{w_{12}\text{Puissance}_1^* + \ldots + w_{16}\text{Largeur}_1^*}{w_{12}^2 + \ldots + w_{16}^2} = -1.80941$$

	Honda Civic
CYLINDREE	?
PUISSANCE	-0.61009
VITESSE	-0.32011
POIDS	-1.10172
LONGUEUR	-1.23196
LARGEUR	-0.32679

	w[1]
CYLINDREE	0.450655
PUISSANCE	0.508011
VITESSE	0.361203
POIDS	0.407744
LONGUEUR	0.376267
LARGEUR	0.317074

Prédiction du Prix :

Prix calculé
= 125 512 + 0.4046*(-1.809)*57 503
= 83 417 FF

Prédiction du prix de la FERRARI (PS)
(Problème : certains X sont manquants)

	Mean	Std.dev	Ferrari	Ferrari (c-r)
CYLINDREE	1887.7	459.31	4943	6.65
PUISSANCE	112.05	36.1422	428	8.74
VITESSE	181.95	24.8352	310	5.16
POIDS	1112.75	238.49	1517	1.70
LONGUEUR	421.7	42.7774	449	0.64
LARGEUR	168.4	7.34417	?	
PRIX	125513	57503.6		

59/200

Prédiction du Prix de la Ferrari

Régression du Prix sur t_1 : $\dfrac{\text{Prix} - 125513}{57503.6} \approx 0.4045789 \times t_1$

Calcul de $tPS_{1i}, \ldots, tPS_{mi}$ pour une nouvelle observation x_i :

- Régression de X sur $t_1, \ldots, t_m$: $X = t_1 p_1' + \ldots + t_m p_m' + \text{résidu}_{WS}$ $\Rightarrow$ les p_h
- Régression de x_i sur $p_1, \ldots, p_m$: $x_i = tPS_{1i} p_1 + \ldots + tPS_{mi} p_m + \text{résidu}_{PS}$ calculée sur les données disponibles; d'où le calcul des tPS_{hi}
- On cherche les tPS_{hi} minimisant la distance entre x_i et le modèle.

Prédiction du prix de la FERRARI (m = 1)

- tPS_1(Ferrari) = 11.376 estimation de $t_{1,25}$
- On utilise tPS1 à la place de t_1

$\Rightarrow$ Prédiction du Prix = 390 172 FF

60/200

Prédiction du Prix de la Ferrari : calcul de tPS_1 (Ferrari)

	P_1
Cylindrée	0.48
Puissance	0.45
Vitesse	0.37
Poids	0.39
Longueur	0.39
Largeur	0.36

$$x_{Ferrari} = \begin{pmatrix} 6.65 \\ 8.74 \\ 5.16 \\ 1.70 \\ 0.64 \\ ? \end{pmatrix} \approx tPS_1(Ferrari) \times \begin{bmatrix} 0.48 \\ 0.45 \\ 0.37 \\ 0.39 \\ 0.39 \\ 0.36 \end{bmatrix} \Rightarrow \boxed{tPS_1(Ferrari) = 11.376}$$

61/200

Régression PLS1 : Cas UOP Guided Wave
Problème : 226 variables X et 26 observations

Les données :

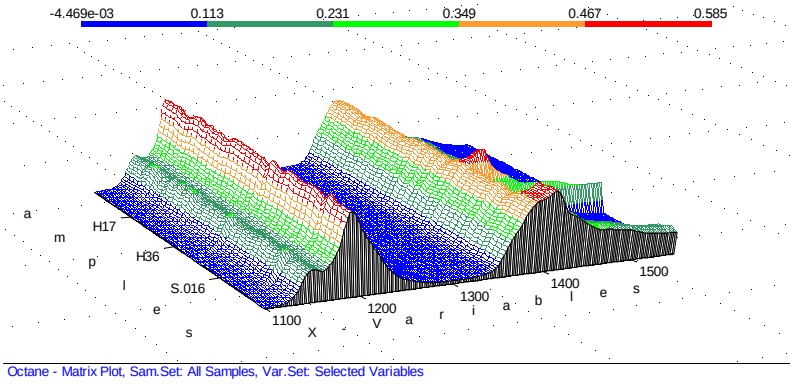
- Y = indice d'octane
- $X_1, X_2, \ldots, X_{226}$:
  valeurs d'absorbance à différentes longueurs d'onde
- *Données de calibration :*

  26 échantillons d'essence (dont 2 avec alcool)
- *Données de validation :*

  13 échantillons d'essence (dont 4 avec alcool)

62/200

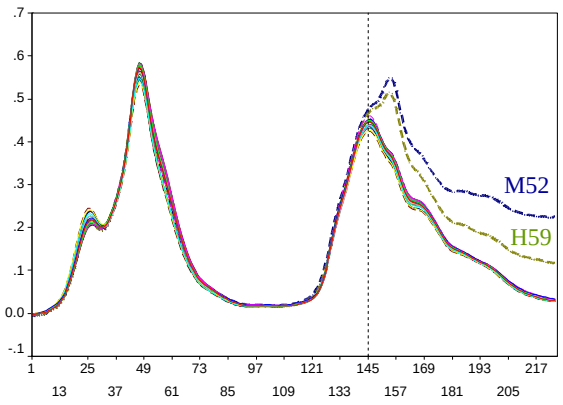
Cas UOP Guided Wave
Visualisation des X



Octane - Matrix Plot, Sam.Set: All Samples, Var.Set: Selected Variables

63/200

Cas UOP Guided Wave
Visualisation des X : Données de calibration



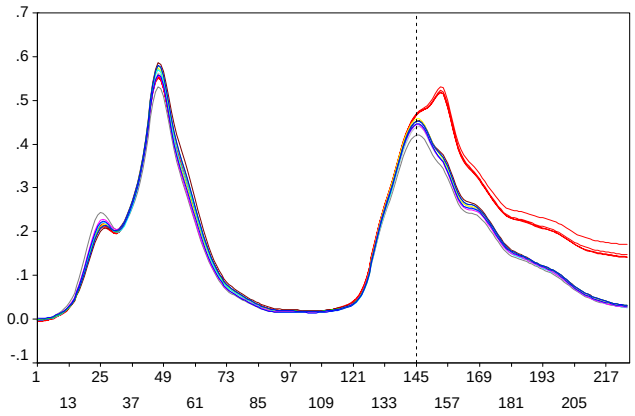
Sequence number

Les échantillons M52 et H59 contiennent de l 'alcool

64/200

Cas UOP Guided Wave

Visualisation des X : Données de validation



Numéro de la longueur d'onde

Les échantillons avec alcool sont en rouge

65/200

Régression PLS1 : les résultats

- Données de spectroscopie

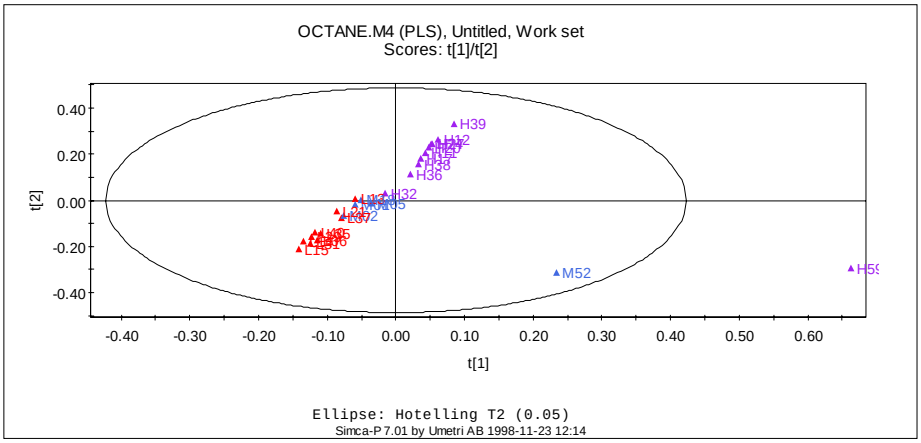
  Les données sont centrées, mais non réduites

- Validation croisée :

  3 composantes PLS

66/200

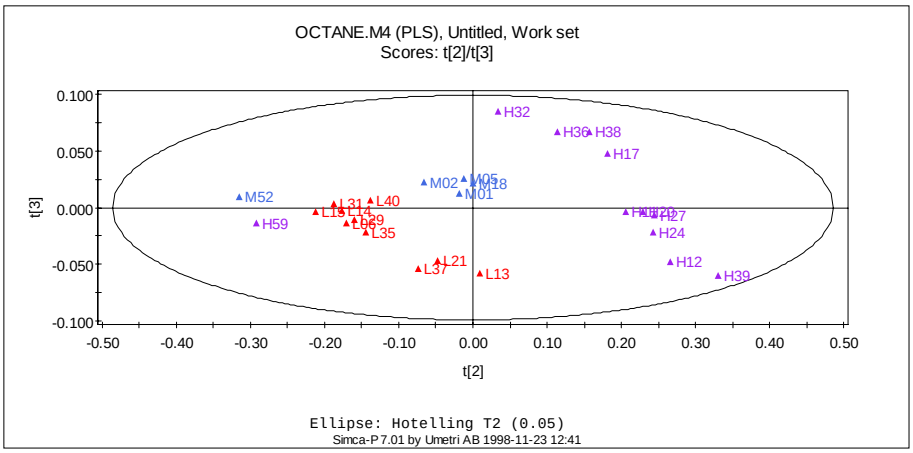
UOP Guided Wave : Les composantes PLS



- Indice d'octane : L = Low, M = Medium, H = High
- Les échantillons M52 et H59 contiennent de l'alcool

67/200

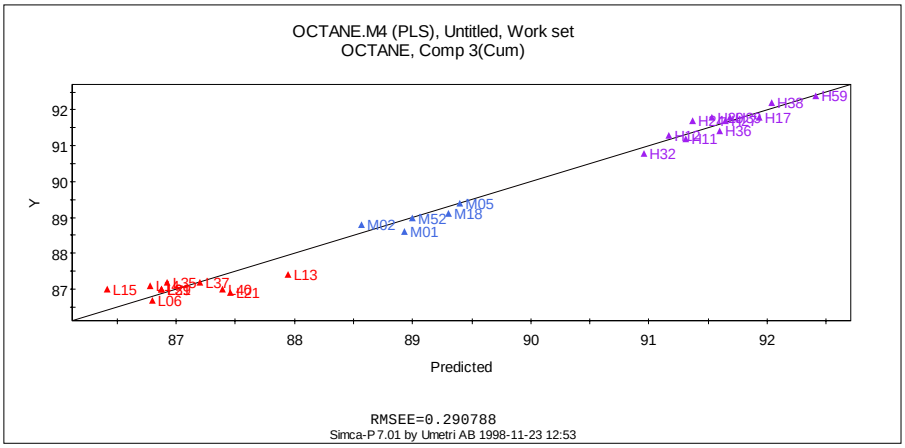
UOP Guided Wave : les composantes PLS



Indice d'octane : L = Low, M = Medium, H = High

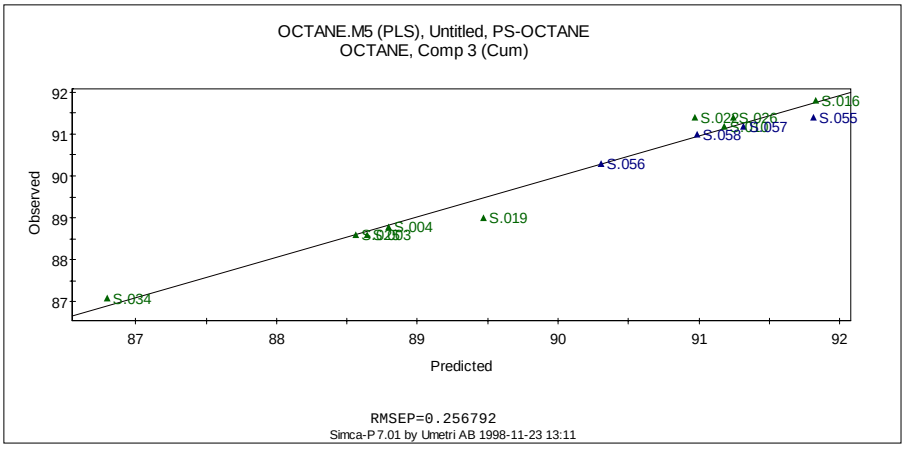
68/200

Cas UOP Guided Wave : Prévision
Données de calibration



69/200

Cas UOP Guided Wave : Prévision
Données de validation



Présence d'alcool : OUI / NON

70/200

Orthogonal Signal Correction

- Filtrage des X pour diminuer la partie des X non corrélée à Y : E=XA
- On recherche une décomposition de X de la forme

$$X = t_{osc,1}p_1' + \ldots + t_{osc,m}p_m' + E$$

avec :

(1) Les composantes $t_{osc,j} = Xw_j$ sont orthogonales.

(2) Les composantes $t_{osc,j}$ sont (à peu près) orthogonales aux Y.

(3) Les p_m et le résidu E sont obtenus par régression multiple de X sur $t_{osc,1},\ldots, t_{osc,m}$.

(4) $D = t_{osc,1}p_1' + \ldots + t_{osc,m}p_m'$ représente la partie des X (à peu près) non corrélée à Y

(5) $E = X - D = X(I - w_1p_1' + \ldots + w_mp_m')$ représente un filtrage des X.

- On effectue la régression PLS de Y sur E.

71/200

Recherche des t_h

(Procédure de Wold et al.)

(1) On réalise une ACP des X
==> première composante principale $t_{osc,1,initial}$

(2) On fait une régression de $t_{osc,1,\ initial}$ sur Y : $t_{osc,1,\ initial} = Yb + t_1^*$

(3) On fait une régression PLS de t_1^* ($\perp$ à Y) sur X avec toutes les composantes ==> $t_{osc,1,\ new} = Xw_1$.
La composante $t_{osc,1,new}$ est peu corrélée à Y et explique bien X.

(4) On itère (2) et (3) jusqu'à convergence ==> $t_{osc,1}$.

(5) On régresse X sur $t_{osc,1}$: $X = t_{osc,1}p_1' + E_1$

(6) Pour obtenir $t_{osc,2}$ on recommence la procédure en remplaçant X par E_1. Et ainsi de suite pour les autres composantes.

72/200

Choix du nombre de composantes $t_{osc,h}$

(1) On déduit de $X = t_{osc,1}p_1' + E_1$ la décomposition

$$\sum_j Var(X_j) = Var(t_{osc,1})\sum_j p_{1j}^2 + \sum_j Var(E_{1j})$$

(2) On mesure la part de X restituée par E_1 par

$$\sum_j Var(E_{1j}) / \sum_j Var(X_j)$$

(3) On conserve $t_{osc,1}$ si

- $t_{osc,1}$ suffisamment orthogonal à Y
- $t_{osc,1}$ explique suffisamment X : (Règle de Wold)

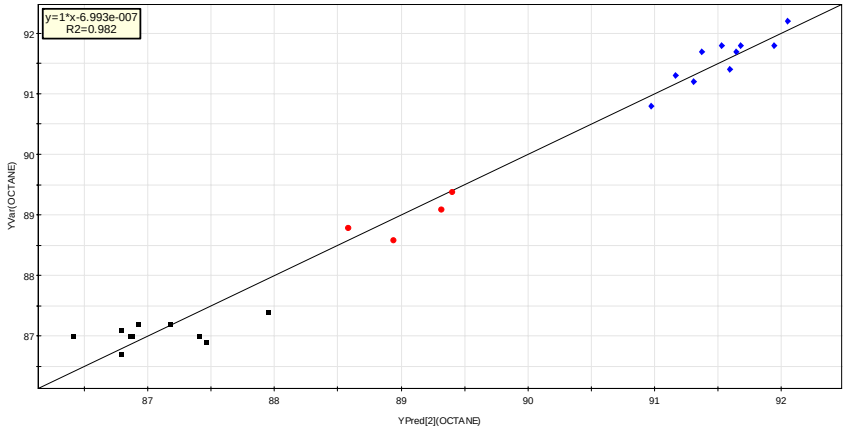
$$\text{"Eigenvalue"} = Var(t_{osc,1})\left[\sum_j p_{1j}^2\right] / \frac{1}{Min(n,p)}\sum_j Var(X_j) \quad > \quad 1 \text{ ou } 2$$

valeur propre moyenne de l'ACP

Et de même pour les autres composantes.

Application à la prédiction de l'indice d'octane

Régression PLS sur les données de calibration non filtrées sans les deux échantillons avec alcool

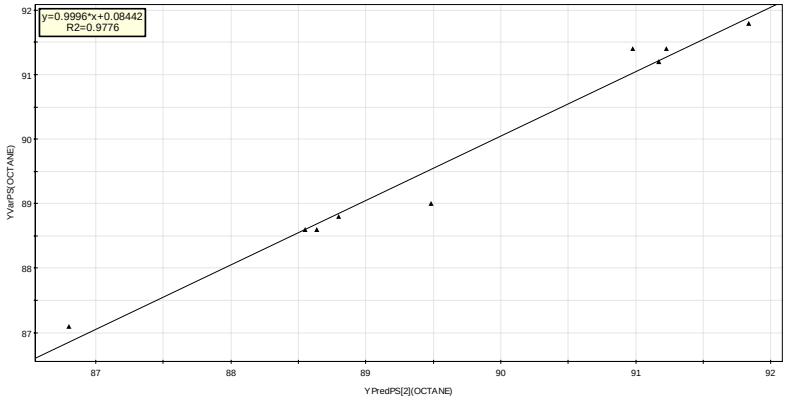


RMSEE = 0.299713

)0

Application à la prédiction de l'indice d'octane

Prédiction sur les données de validation sans les quatre échantillons avec alcool

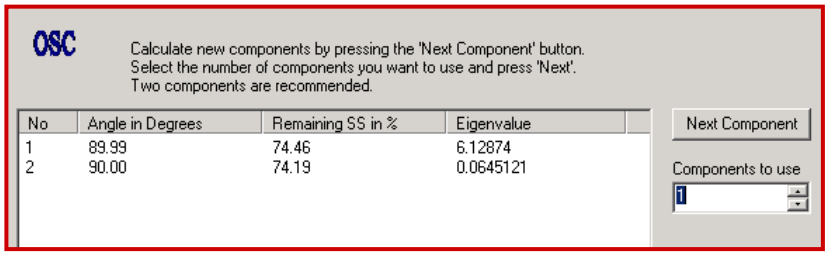


RMSEP = 0.24295

75/200

Application à la prédiction de l'indice d'octane

Filtrage des X par OSC sur l'échantillon de calibration sans les deux échantillons avec alcool



On retient la seule composante t_1.

76/200

Application à la prédiction de l'indice d'octane

Valeurs de $t_{osc,1}$

Obs ID (Primary)	OSC.t[1]
M01	0.0003697
M02	0.00474571
M05	0.00364934
L06	0.00095048
H11	-0.00256955
H12	-0.00573855
L13	-0.00841451
L14	0.00269339
L15	0.00381814
H17	0.00319899
M18	0.00064989
H20	-0.00134013
L21	-0.00678636
H24	-0.0034062
H27	-0.00306896
L29	0.00081317
L31	0.00305239
H32	0.0081986
L35	0.00015011
H36	0.00530462
L37	-0.00315448
H38	0.00551934
H39	-0.0087859
L40	0.00015085

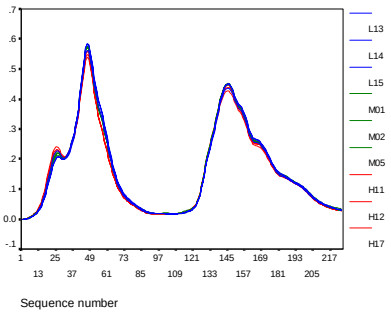
En résumé

- On régresse X sur t_1, partie de X orthogonale à Y.
- D'où le résidu E_1, données filtrées par OSC.
- Puis on réalise la régression PLS de Y sur le résidu E_1

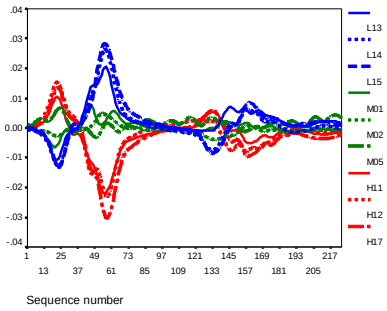
77/200

Application à la prédiction de l'indice d'octane

Comparaison entre les données brutes et les données filtrées par OSC



Données brutes

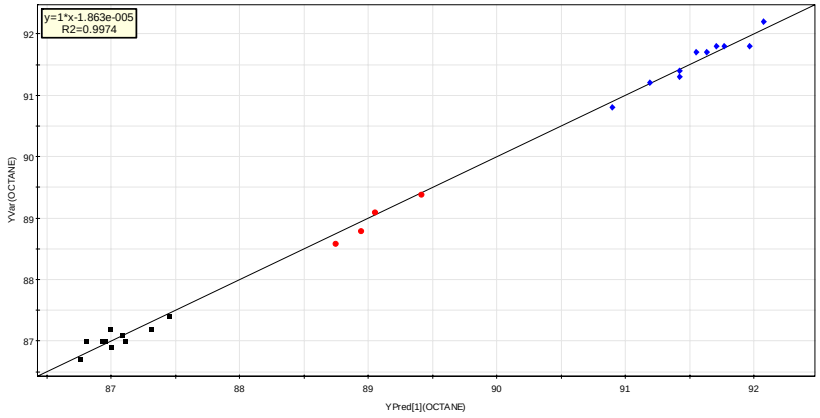


Données filtrées

78/200

Application à la prédiction de l'indice d'octane

Régression PLS sur les données de calibration (sans les deux échantillons avec alcool) filtrées par OSC avec une seule composante

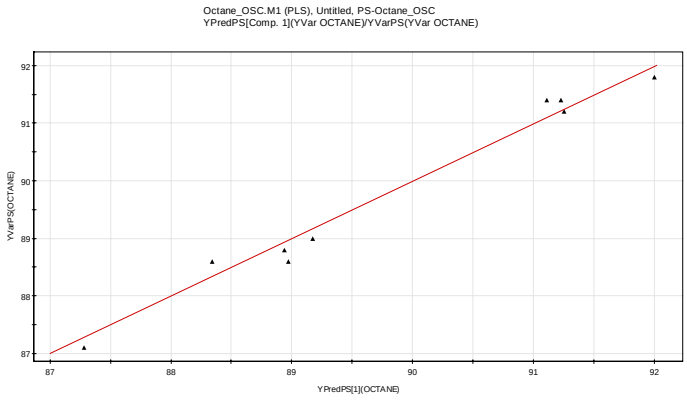


RMSEE = 0.112036

200

Application à la prédiction de l'indice d'octane

Prédiction de l'indice d'octane sur les données de validation sans les quatre échantillons avec alcool



RMSEP = 0.222613

SIMCA-P 10.5 - 15/06/2004 10:03:34

80/200

Les méthodes PLS

II.2 Régression PLS2

81/200

La régression PLS2

- Relier un bloc de variables à expliquer Y à un bloc de variables explicatives X.
- Possibilité de données manquantes.
- Il peut y avoir beaucoup plus de variables X que d'observations.
- Il peut y avoir beaucoup plus de variables Y que d'observations.

82/200

La régression PLS2 : une idée de l'algorithme

Etape 1 : Recherche de m composantes orthogonales $t_h = Xa_h$ et m composantes $u_h = Yb_h$, bien corrélées entre elles et explicatives de leur propre groupe.

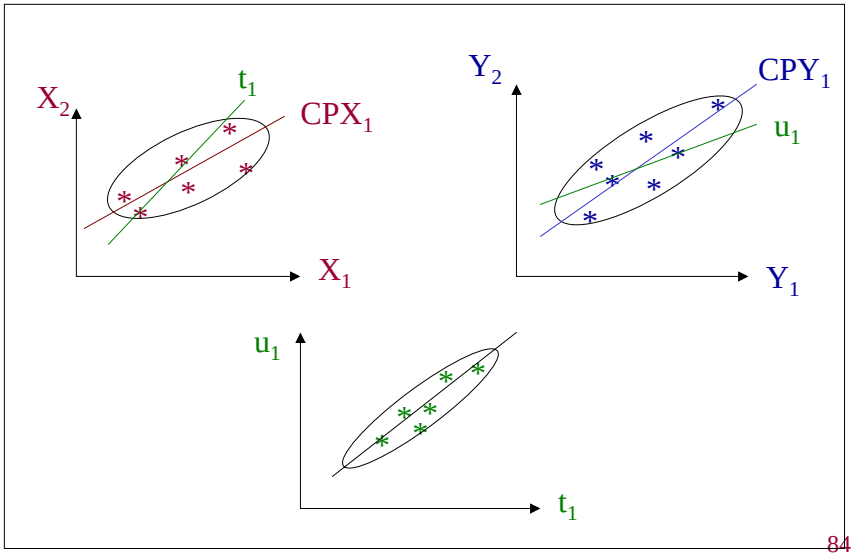
Le nombre m est obtenu par validation croisée.

Etape 2 : Régression de Y sur les composantes t_h .

Etape 3 : Expression de la régression en fonction de X.

83/200

Objectif de l'étape 1 de la régression PLS2



84/200

La régression PLS2 : une idée de l'étape 1 lorsqu'il n'y a pas de données manquantes

Pour chaque h = 1 à m, on recherche des composantes $t_h = Xa_h$ et $u_h = Yb_h$ maximisant le critère

$$Cov\,(Xa_h\,,\,Yb_h\,)$$

sous des contraintes de norme et d'orthogonalité entre t_h et les composantes précédentes t_1 ,…, t_{h-1}.

85/200

Interprétation du critère de Tucker

De $Cov^2(Xa_h\,,\,Yb_h\,)$

$= Cor^2(Xa_h\,,\,Yb_h\,)^* Var(Xa_h)^*Var(Yb_h)$

on déduit que la régression PLS réalise un compromis entre l'analyse canonique de X et Y, une ACP de X, et une ACP « oblique » de Y.

86/200

Variable Importance in the Prediction (VIP)

- Composantes PLS : $t_h = X_{h-1} b_h$, avec $||b_h|| = 1$
- Importance de la variable x_j (j=1, p) pour la prédiction des y_k (k=1, q) dans un modèle à *m* composantes :

$$VIP_{mj} = \sqrt{\frac{p}{\sum_{h=1}^{m}\sum_{k=1}^{q} R^2(y_k;t_h)} \sum_{h=1}^{m}[\sum_{k=1}^{q} R^2(y_k,t_h)]b_{hj}^2}$$

Pouvoir prédictif de X_j

- Moyenne des carrés des VIP = 1
- Variable importante pour la prévision si VIP > 0.8

87/200

Régression PLS2

Exemple : Dégustation de thé

Les données

Obs	Température	Sucré	Force	Citron	Sujet 1	…	Sujet 6
1	1	1	1	1	4		5
2	1	2	2	1	2		8
3	1	3	3	2	6		6
⋮							
11	1	2	1	1	1		14
⋮							
18	3	3	1	2	12		15

Température	Sucré	Force	Citron
1 = Chaud 2 = Tiède 3 = Glacé	1 = Pas de sucre 2 = 1 sucre 3 = 2 sucres	1 = Fort 2 = Moyen 3 = Faible	1 = Avec 2 = Sans

88/200

Cas Dégustation de thé

- Bloc X
  Les 11 variables indicatrices des modalités de Température, Sucré, Force et Citron

- Bloc Y
  Les classements des sujets sont inversés
  (Le 1 devient 18, le 18 devient 1)

89/200

Cas Dégustation de thé

Résultats de la régression PLS

- Validation croisée :

  3 composantes : $t_h = Xw_h^*$ et $u_h = Yc_h$

- Équation de régression de Y_k sur $t_1, \ldots, t_h$:

  $Y_k = c_{1k}t_1 + c_{2k}t_2 + c_{3k}t_3 + \text{résidu}$

- Les variables X et Y sont représentées à l'aide des vecteurs w_h^* et c_h.

90/200

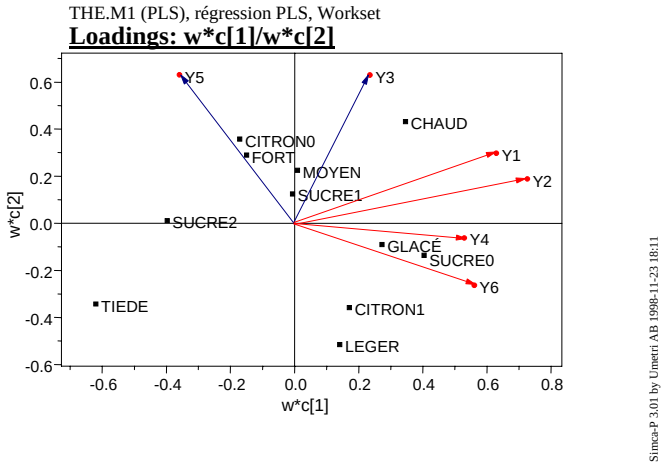
Résultats de la régression PLS
Les vecteurs w*c

	w*c[1]	w*c[2]	w*c[3]
CHAUD	0.346	0.432	-0.393
TIEDE	-0.620	-0.342	0.245
GLACE	0.273	-0.090	0.148
SUCRE0	0.404	-0.136	0.538
SUCRE1	-0.006	0.124	-0.031
SUCRE2	-0.397	0.012	-0.507
FORT	-0.150	0.289	0.125
MOYEN	0.009	0.225	-0.069
LEGER	0.140	-0.515	-0.056
CITRON1	0.171	-0.358	-0.316
CITRON0	-0.171	0.358	0.316
Y1	0.629	0.298	-0.298
Y2	0.726	0.188	0.146
Y3	0.235	0.630	-0.203
Y4	0.530	-0.062	0.538
Y5	-0.359	0.631	0.216
Y6	0.561	-0.263	-0.035

w* : lignes CHAUD à CITRON0

c : lignes Y1 à Y6

Cas Dégustation de thé
Carte des variables

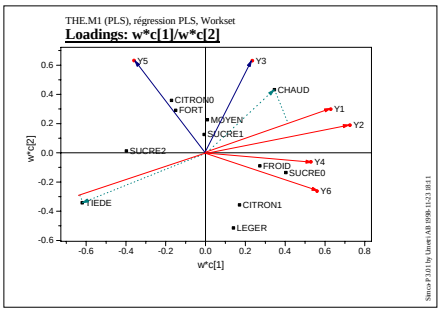


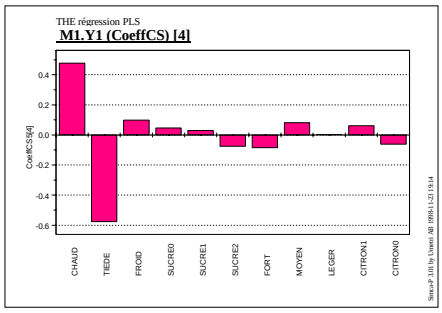
Règle d'interprétation

Les projections des variables X sur les variables Y reflètent le signe et l'ordre de grandeur des coefficients de régression PLS des Y sur X

Cas dégustation de thé

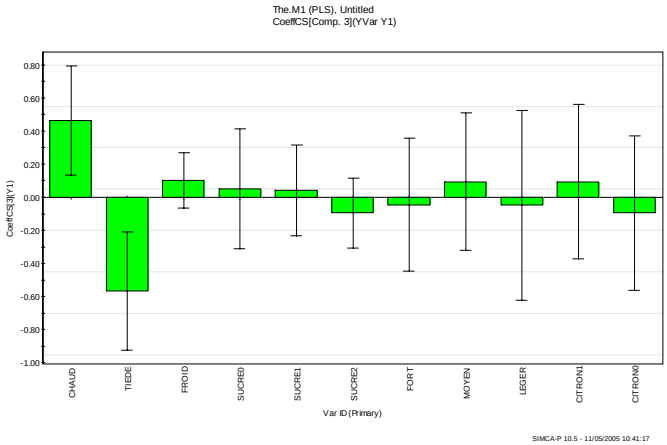
Visualisation de la régression PLS de Y_1 sur X





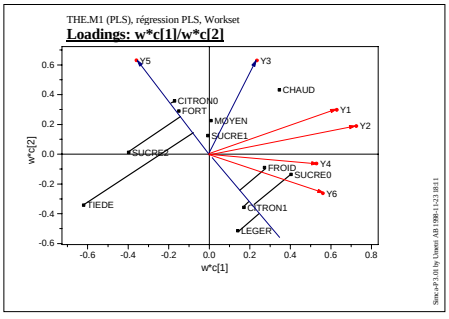
Le juge 1 aime son thé chaud et rejette le thé tiède

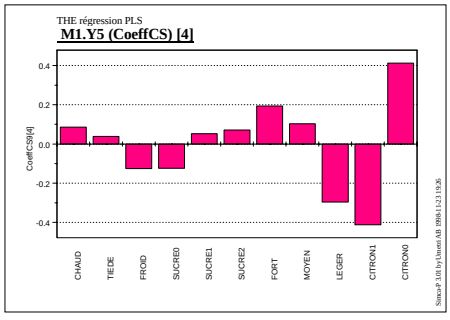
Validation du modèle pour le juge 1



Cas dégustation de thé

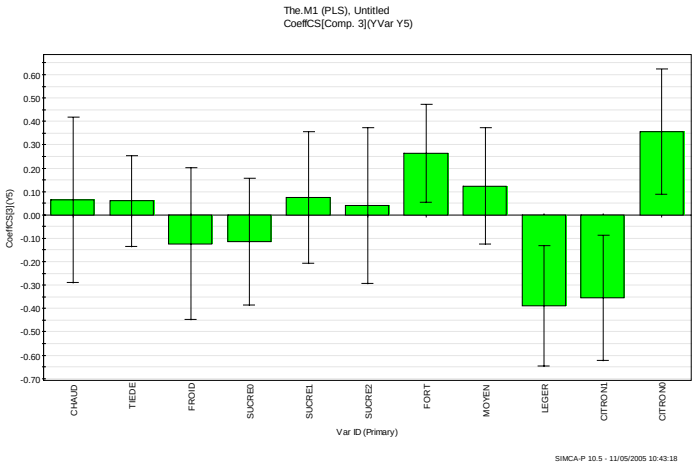
Visualisation de la régression PLS de Y_5 sur X





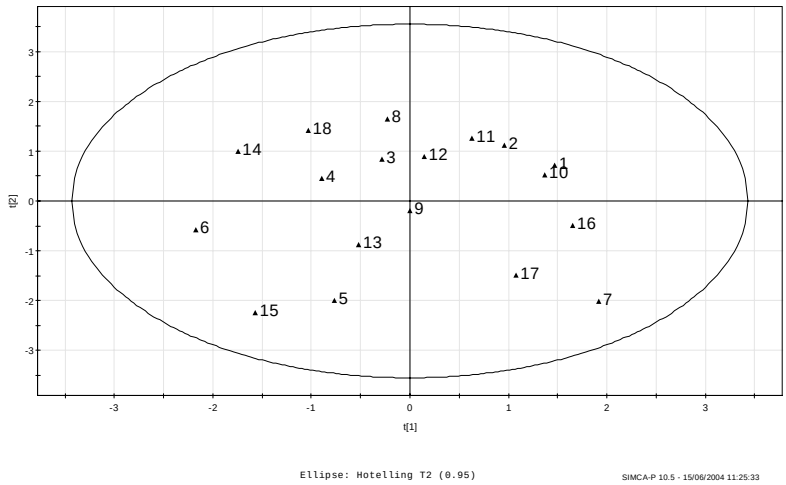
Le juge 5 préfère son thé sans citron, fort;
il est indifférent au thé tiède; il rejette le thé léger, avec du citron.

Validation du modèle pour le juge 5



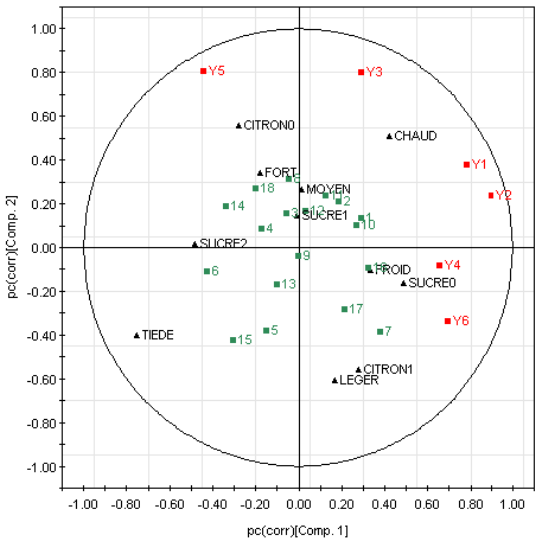
97/200

Carte des produits



98/200

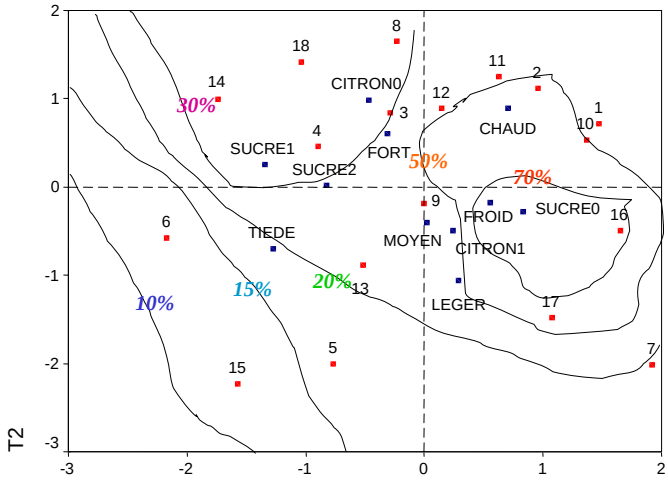
Carte des corrélations : Biplot



99/200

Carte des produits et des caractéristiques

Courbes de niveau du % de juges classant le produit au dessus de leurs moyennes



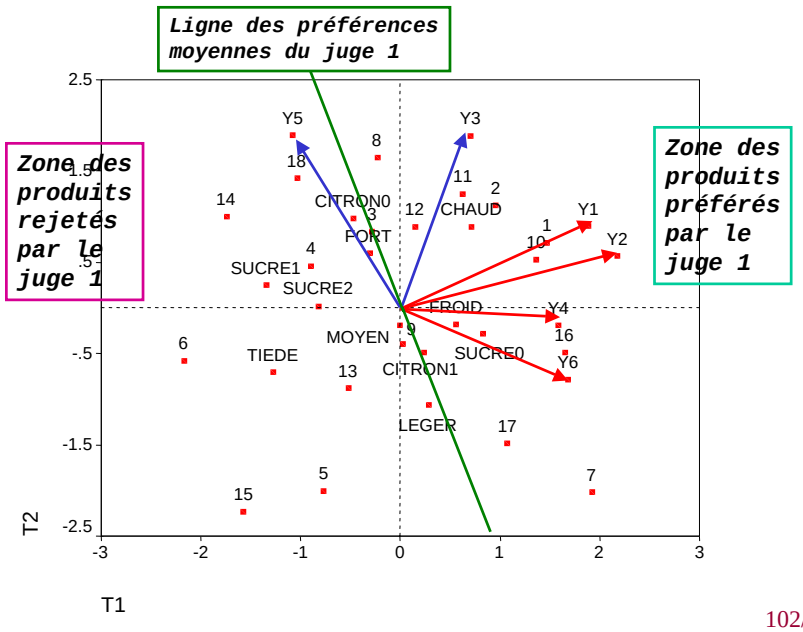
Courbes simulées non calculées sur les données

100/200

Construction des courbes de niveau

- On utilise la carte des produits et des caractéristiques dans le plan des composantes PLS (t_1, t_2).
- On représente dans ce plan les juges Y_k à l'aide des points ($3c_{k1}$, $3c_{k2}$).
- Pour chaque juge k le plan (t_1, t_2) est partagé en deux zones : la zone des « 1 » pour Y_k estimé $= c_{k1}t_1 + c_{k2}t_2 > 0$ et la zone des « 0 » pour Y_k estimé ≤ 0.
- Pour chaque point du plan (t_1, t_2) on détermine le % de « 1 » = % de juges classant le produit avec les caractéristiques correspondant à ce point au dessus de la moyenne.
- On résume tous ces pourcentages par des courbes de niveau.

Carte des produits, des caractéristiques et des juges



Typologie des juges

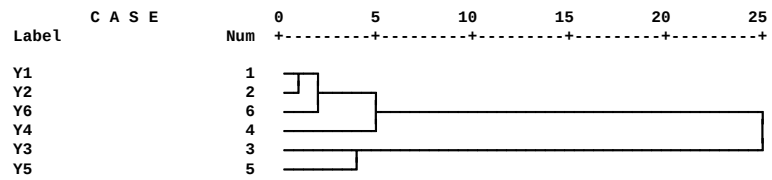
- Construire une typologie des juges en utilisant la part des préférences explicables par les caractéristiques des produits.
- Le vecteur $c_k = (c_{1k},c_{2k},c_{3k})$ représente la part des préférences du juge k expliquée par les caractéristiques des produits.
- On construit la typologie des juges à l'aide des c_k.

Typologie des juges

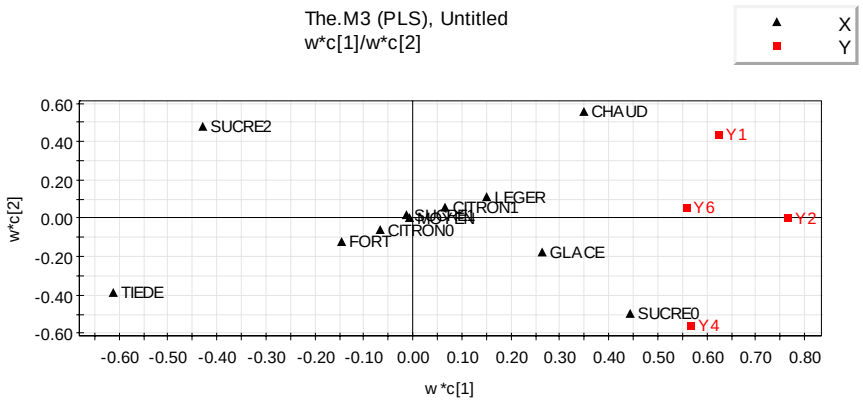
* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Ward Method

Rescaled Distance Cluster Combine

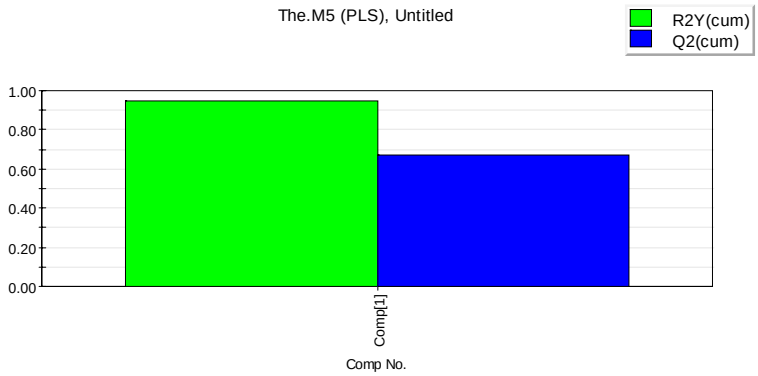


Régression PLS sur le premier groupe



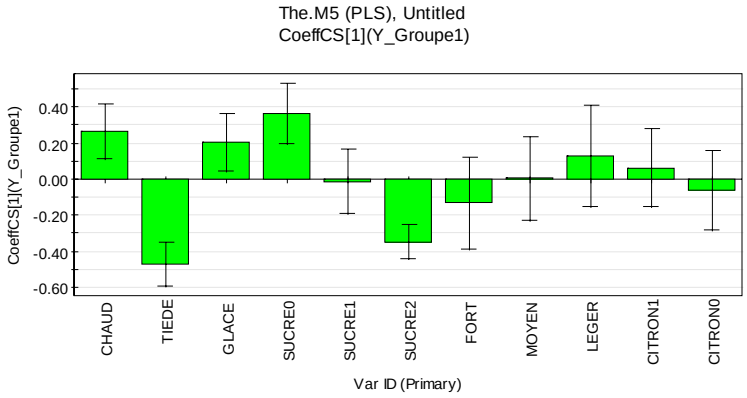
105/200

Régression PLS sur la moyenne des juges du premier groupe

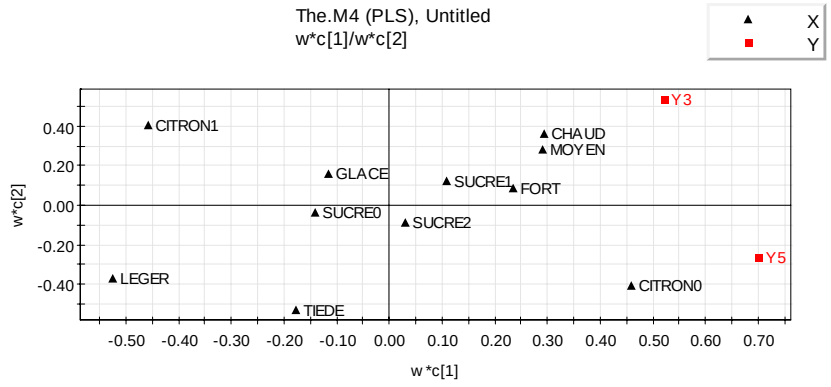


106/200

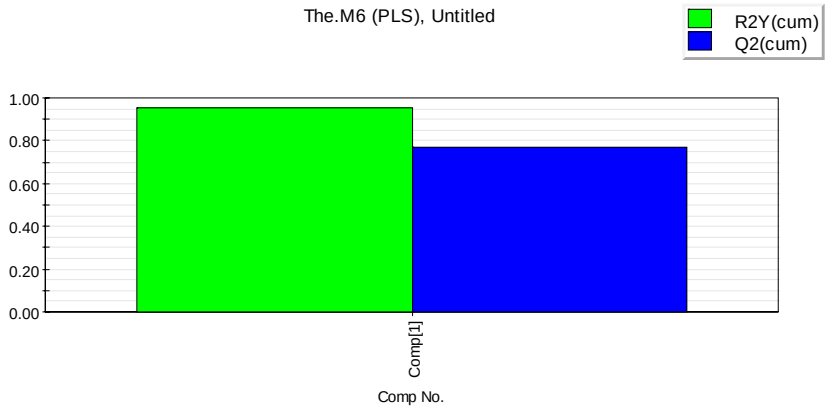
Régression PLS sur la moyenne des juges du premier groupe : modalités significatives



Régression PLS sur le deuxième groupe

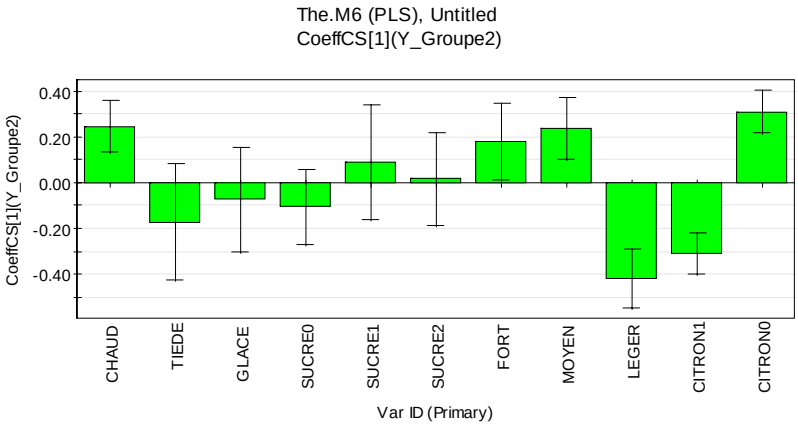


Régression PLS sur la moyenne des juges du deuxième groupe



109/200

Régression PLS sur la moyenne des juges du deuxième groupe : modalités significatives



110/200

Les méthodes PLS

III. Analyse discriminante PLS

Analyse discriminante PLS

- Bloc Y

  La variable qualitative Y est remplacée par l'ensemble des variables indicatrices de ses modalités.

- Bloc X

  Variables numériques ou indicatrices des modalités des variables qualitatives.

- Régression PLS2 de Y sur X

Analyse discriminante PLS : exemple

Jellum E., Bjørnson I., Nesbakken R., Johanson E., Wold S. :

Classification of human cancer cells by means of capillary gas chromatography and pattern recognition analysis.

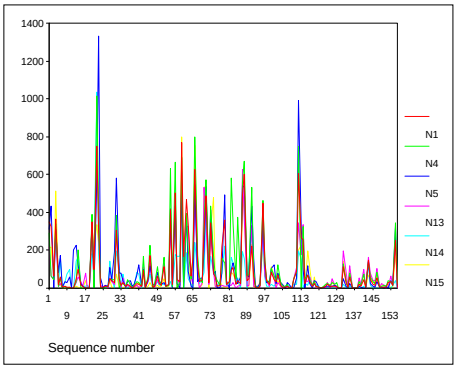
Journal of Chromatography, 1981

Analyse discriminante PLS : exemple

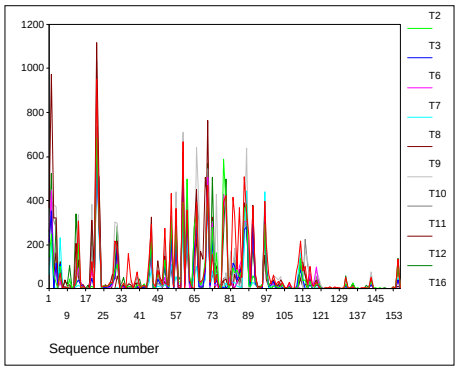
Les données

- 16 biopsies de tumeurs de cerveau humain.
- Chaque tumeur est classée par un médecin anatomo-pathologiste comme bénigne ou maligne.
- Chaque biopsie est analysée par chromatographie en phase gazeuse : on obtient un profil métabolique de la biopsie formé de 156 pics.
- Quelques données manquantes

Analyse discriminante PLS
Profils métaboliques des biopsies



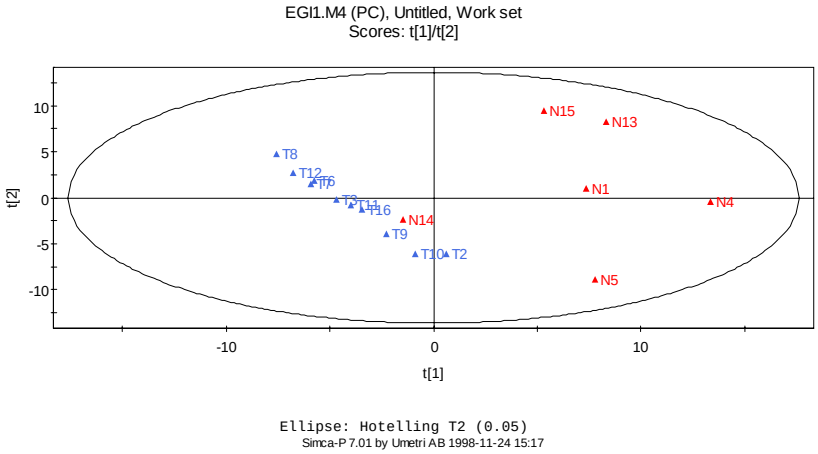
Tumeurs bénignes



Tumeurs malignes

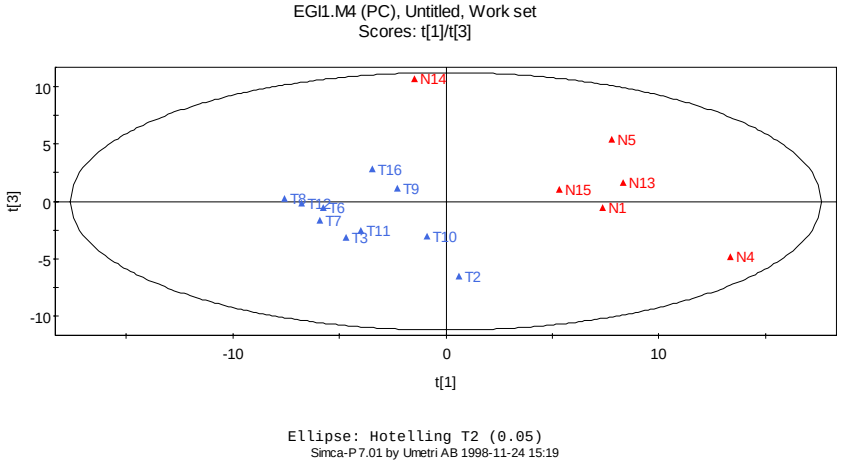
115/200

Analyse en composantes principales des 16 biopsies
Composantes principales 1 et 2



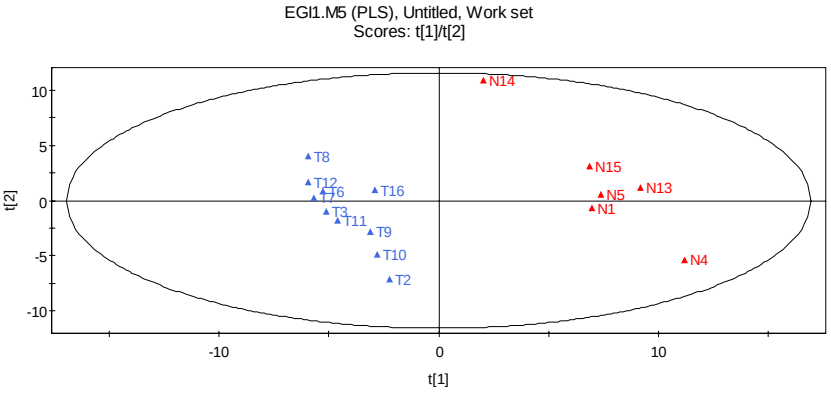
116/200

Analyse en composantes principales des 16 biopsies
Composantes principales 1 et 3



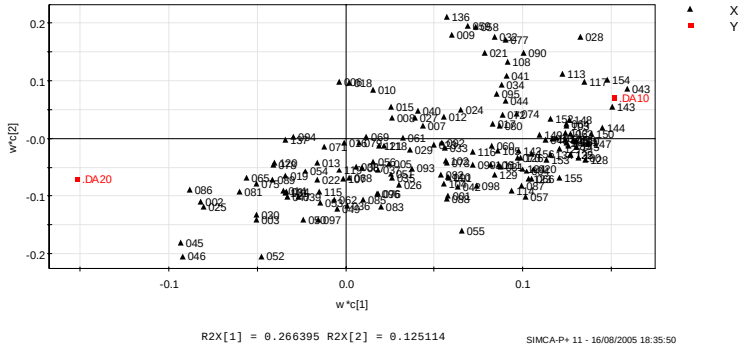
117/200

Analyse discriminante PLS
Composantes PLS 1 et 2



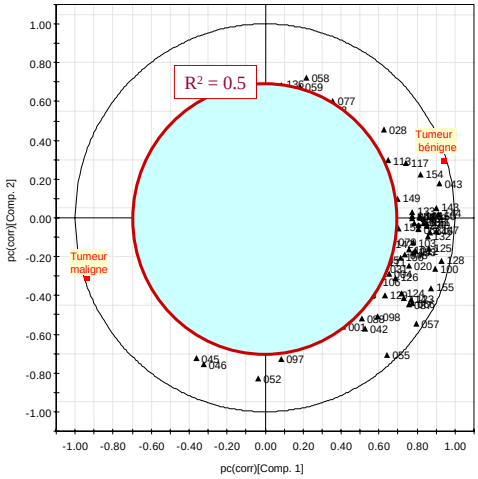
118/200

Graphique des variables w*c

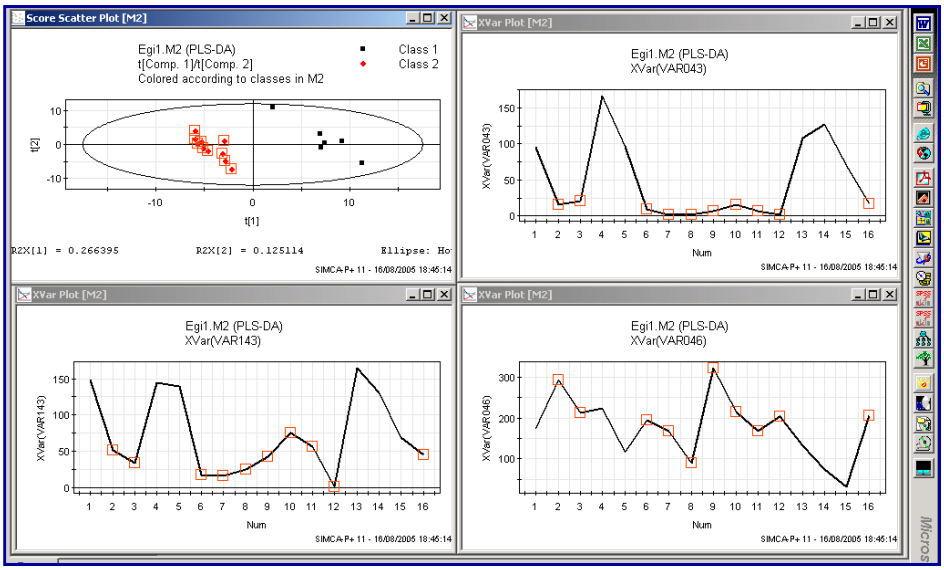


119/200

Analyse discriminante PLS
Carte des variables

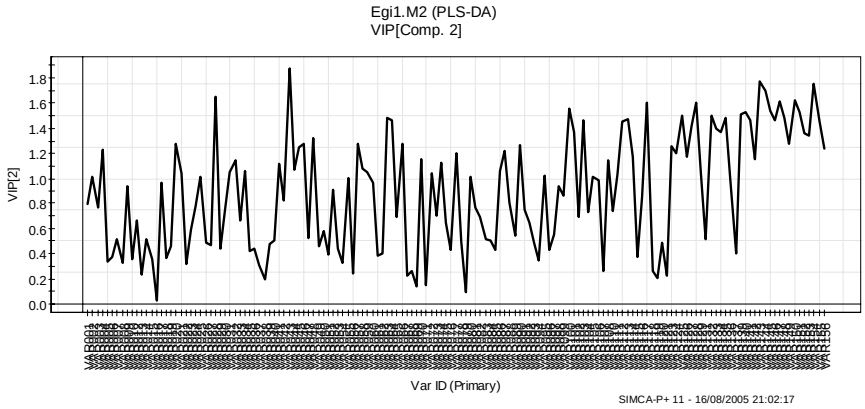


Visualization Plots



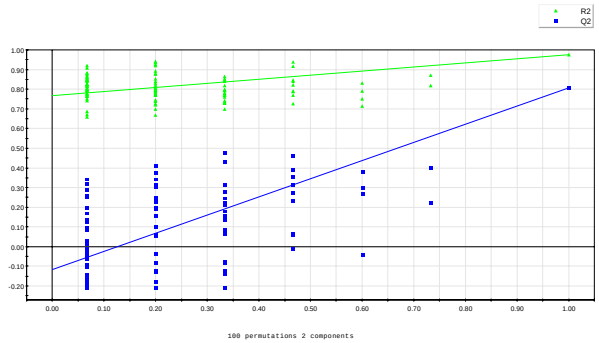
121/200

Graphique des VIP



122/200

Validation globale



100 permutations, 2 composantes

Modèle validé : L'ordonnée à l'origine de la droite $Q^2 < 0$

123/200

Tableau de classification

- $Y_k > .65$ ➔ Classé dans le groupe k (vert)
- $Y_k < .35$ ➔ Non classé dans le groupe k (blanc)
- $.35 < Y_k < .65$ ➔ pas de décision (orange)

	Members	Class 1	Class 2	No class	Class 1 & 2
Class 1	6	6	0	0	0
Class 2	10	0	10	0	0
No class	0	0	0	0	0

124/200

Classification des individus

	1	2	3	4	5
1	Obs ID (Primary)	M2.YVarPS($M2.DA10)	M2.YPredPS[2]($M2.DA10)	M2.YVarPS($M2.DA20)	M2.YPredPS[2]($M2.DA20)
2	N1	1	0.882442	0	0.117558
3	T2	0	-0.0478299	1	1.04783
4	T3	0	-0.0486789	1	1.04868
5	N4	1	1.0338	0	-0.033801
6	N5	1	0.956843	0	0.043157
7	T6	0	0.00467116	1	0.995329
8	T7	0	-0.0498978	1	1.0499
9	T8	0	0.0652757	1	0.934724
10	T9	0	0.0377735	1	0.962227
11	T10	0	-0.0115993	1	1.0116
12	T11	0	-0.0400793	1	1.04008
13	T12	0	-0.0156377	1	1.01564
14	N13	1	1.11554	0	-0.115537
15	N14	1	0.919642	0	0.0803581
16	N15	1	1.01013	0	-0.0101316
17	T16	0	0.189296	1	0.810704

125/200

Les méthodes PLS

IV. SIMCA

Soft Independent Modelling by Class Analogy

126/200

SIMCA
(Soft Independent Modelling of Class Analogy)

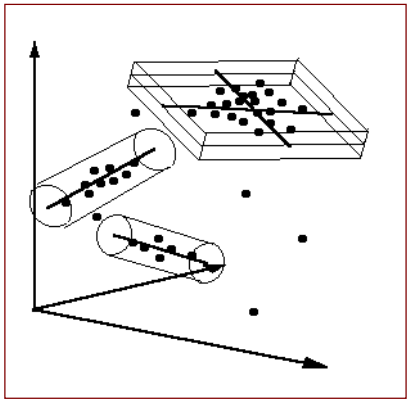
On réalise une analyse en composantes principales sur chaque classe *h* via NIPALS

$\Rightarrow$ - Choix automatique du nombre de composantes.

- Possibilité de données manquantes.

127/200

SIMCA
(Soft Independent Modelling of Class Analogy)



- Calcul de la distance entre chaque individu i et le modèle ACP de la classe *h*.
- Calcul de la « probabilité » d'appartenance de chaque individu à la classe *h*.

128/200

Utilisation de SIMCA sur les 16 biopsies

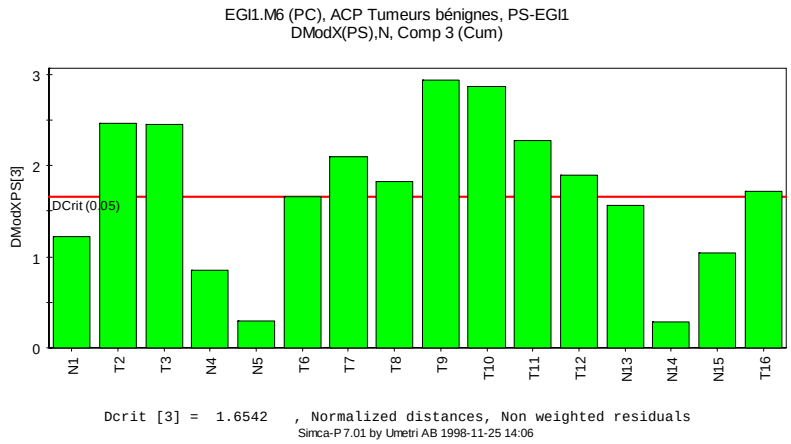
- L ’ACP des tumeurs bénignes conduit à 3 composantes.
- L ’ACP des tumeurs malignes conduit à 4 composantes.
- Le Cooman’s Plot permet de visualiser les distances de chaque biopsie aux deux classes.

129/200

Utilisation de SIMCA sur les 16 biopsies

ACP des tumeurs bénignes :

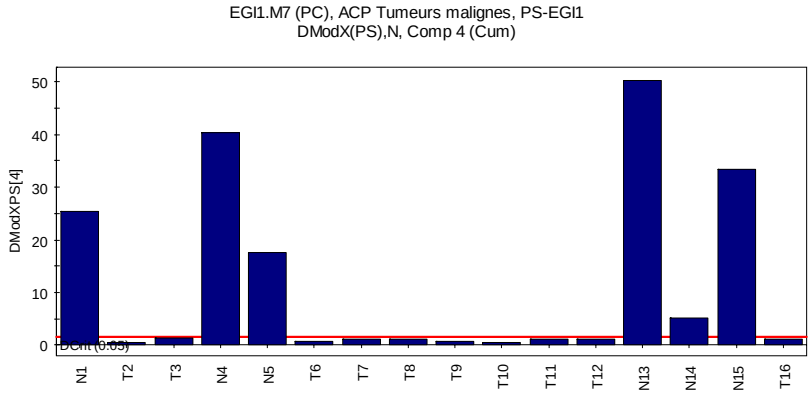
DModX (Classe des tumeurs bénignes)



130/200

Utilisation de SIMCA sur les 16 biopsies

ACP des tumeurs malignes :
DModX (Classe des tumeurs malignes)




```
Dcrit [4] =  1.4708   , Normalized distances, Non weighted residuals
```

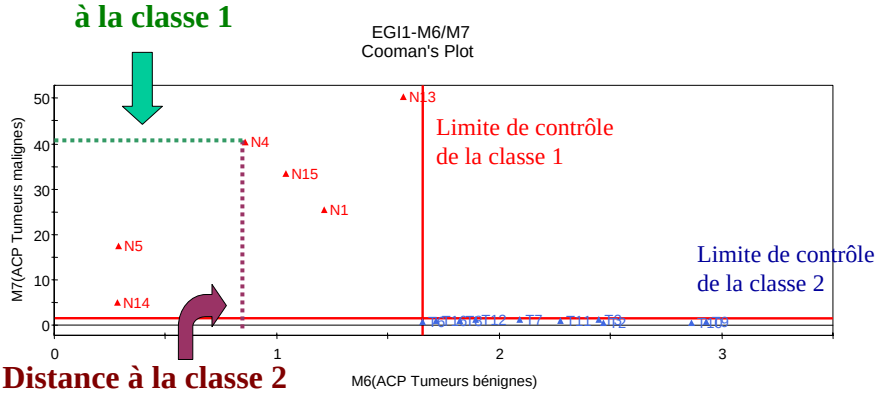
Simca-P 7.01 by Umetri AB 1998-11-25 14:12

131/200

Utilisation de SIMCA sur les 16 biopsies

Cooman's Plot

Distance à la classe 1



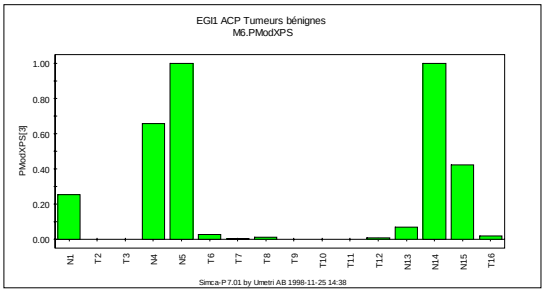
Distance à la classe 2

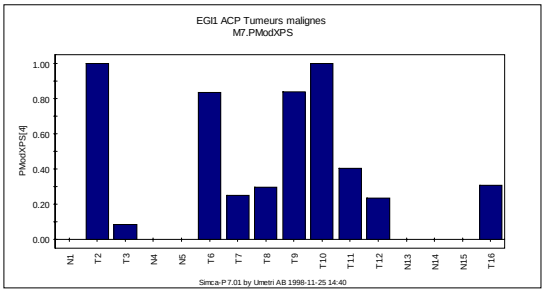
```
M6-crit [3] =  1.6542   , Normalized distances, Non weighted residual
M7-crit [4] =  1.4708   , Normalized distances, Non weighted residual
```

Simca-P 7.01 by Umetri AB 1998-11-25 14:20

132/200

Les « probabilités » d ’appartenance aux classes :
« Probabilité (biopsie/classe) »



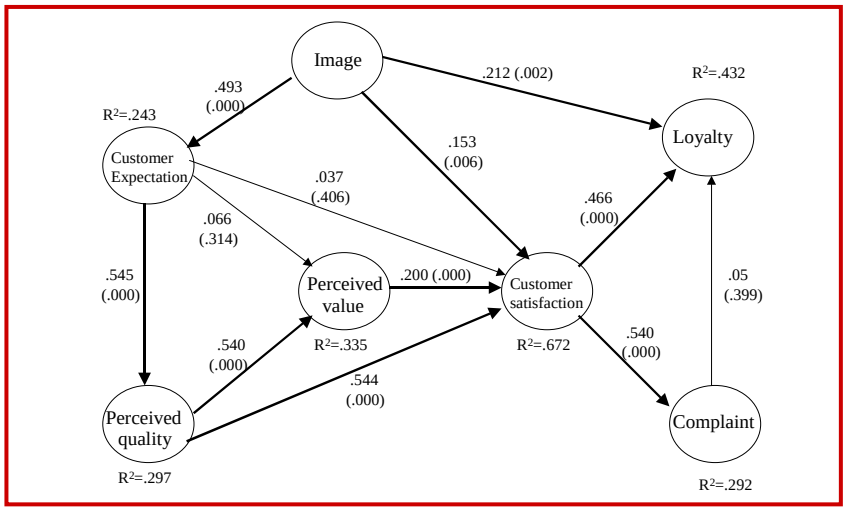


Les méthodes PLS

V. PLS Path Modelling
(Approche PLS)

Modélisation de relations structurelles sur variables latentes

ECSI Path model for a " Mobile phone provider"



135/200

Modélisation de relations structurelles L'approche PLS de Herman WOLD

- Etude d'un système de relations linéaires entre variables latentes (non observables).
- Chaque variable latente est décrite par des variables manifestes (observables).
- Les données sont quantitatives ou qualitatives (pas d'hypothèse de normalité).
- Le nombre d'observations peut être limité par rapport au nombre de variables.

136/200

Inégalité économique et instabilité politique
(Données de Russett, 1964)

Inégalité économique

Inégalité agricole

GINI : Inégalité dans la répartition des terres

FARM : % fermiers possédant la moitié des terres (> 50%)

RENT : % fermiers locataires

Développement industriel

GNPR : PNB par habitant ($ 1955)

LABO : % d 'actifs dans l'agriculture

Instabilité politique

INST : Instabilité de l 'exécutif (45-61)

ECKS : Nb de conflits violents entre communautés (46-61)

DEAT : Nb de morts dans des manifestations (50-62)

D-STAB : Démocratie stable

D-INS : Démocratie instable

DICT : Dictature

137/200

Inégalité économique et instabilité politique
(Données de Russett, 1964)

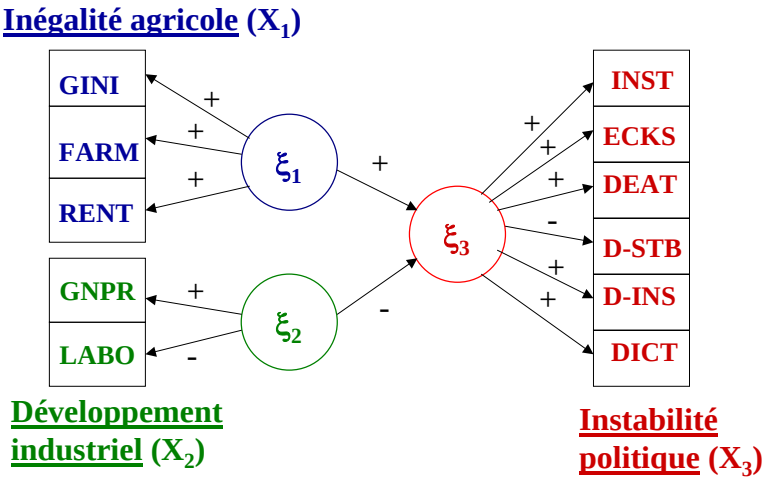
47 pays

	Gini	Farm	Rent	Gnpr	Labo	Inst	Ecks	Deat	régime
Argentine	86.3	98.2	32.9	374	25	13.6	57	217	2
Australie	92.9	99.6	*	1215	14	11.3	0	0	1
Autriche	74.0	97.4	10.7	532	32	12.8	4	0	2
⋮									
France	58.3	86.1	26.0	1046	26	16.3	46	1	2
⋮									
Yougoslavie	43.7	79.8	0.0	297	67	0.0	9	0	3

1 = Démocratie stable
2 = Démocratie instable
3 = Dictature

138/200

Inégalité économique et instabilité politique



Inégalité économique et instabilité politique
Le modèle

- Chaque variable manifeste X_{jh} s 'écrit :

$$X_{jh} = \pi_{jh}\xi_h + \varepsilon_{jh}$$

- Il existe une relation structurelle entre les variables latentes :

Instabilité politique (ξ_3)
= β_1×Inégalité agricole (ξ_1) + β_2× Dév. industriel (ξ_2)
+ résidu

Estimation des variables latentes par la méthode PLS

(1) Estimation externe Y_h de ξ_h :

$$Y_h = X_h w_h$$

(2) Estimation interne Z_h de ξ_h :

$$Z_h = \sum_{\substack{j \neq h \\ \xi_j \text{ reliée à } \xi_h}} [\text{signe}(\text{cor}(\xi_j, \xi_h))] Y_j$$

(3) Calcul de w_h :

$$w_{hj} = \text{cor}(Z_h , X_{hj})$$

141/200

Inégalité économique et instabilité politique
Estimation des variables latentes par la méthode PLS

(1) Estimation externe

$Y_1 = X_1 w_1$

$Y_2 = X_2 w_2$

$Y_3 = X_3 w_3$

(2) Estimation interne

$Z_1 = Y_3$

$Z_2 = -Y_3$

$Z_3 = Y_1 - Y_2$

(3) Calcul des w_h

$w_{1j} = \text{cor}(X_{1j} , Z_1)$

$w_{2j} = \text{cor}(X_{2j} , Z_2)$

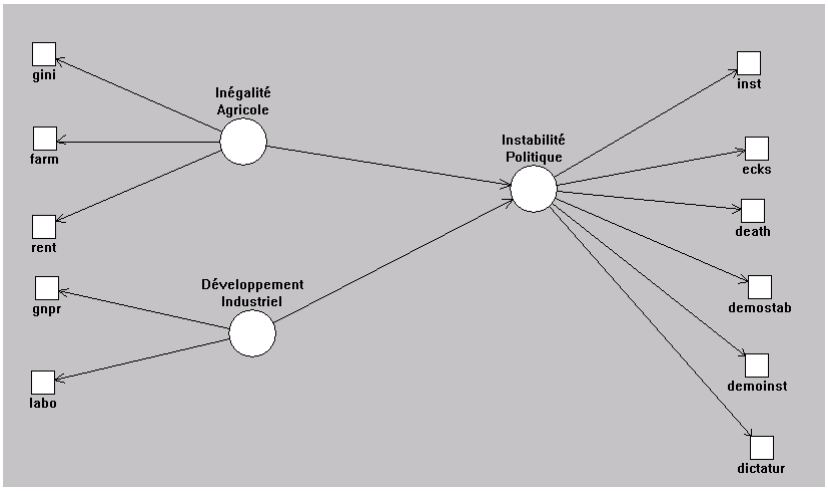
$w_{3j} = \text{cor}(X_{3j} , Z_3)$

L'algorithme

- On part de w_1, w_2, w_3 arbitraires.
- On obtient de nouveaux w_h en utilisant (1) à (3).
- On itère jusqu'à convergence.

142/200

Utilisation de PLS-Graph de Wynne Chin



143/200

Résultats

```
Outer Model
================================
Variable      Weight   Loading (Corrélation)
--------------------------------
   ineg_agr  outward
gini            .4567     .9745
farm            .5125     .9857
rent            .1018     .5156
--------------------------------
   dev_indu  outward
gnpr            .5113     .9501
labo           -.5384    -.9551
--------------------------------
   inst_pol  outward
inst            .1187     .3676
ecks            .2855     .8241
deat            .2977     .7910
demostab       -.3271    -.8635
demoinst        .0370     .1037
dictatur        .2758     .7227
================================
```

144/200

Résultats

```
Eta  .. Latent variables
=======================================
            ineg_agr  dev_indu  inst_pol
---------------------------------------
arg             .964      .238      .755
aus            1.204     1.371    -1.617
aut             .397      .253     -.480
bel            -.812     1.530     -.846
bol            1.115    -1.584     1.505
bré             .778     -.654      .302
.
.
.
tai            -.009     -.898     -.068
ru              .134     2.059    -1.046
eu              .193     2.016     -.942
uru             .699      .179    -1.298
ven            1.149      .252     1.135
rfa            -.212     1.104     -.494
you           -2.189     -.654      .125
=======================================
```

145/200

Les résultats de PLS

Estimation des variables latentes

	Y_1	Y_2	Y_3
Argentine	0.96	0.24	0.75
Australie	1.21	1.37	-1.62
Autriche	0.40	0.25	-0.48
:			
France	-0.89	0.80	0.56
:			
Yougoslavie	-2.18	-0.65	0.13

Régression multiple de Y_3 sur Y_1 et Y_2

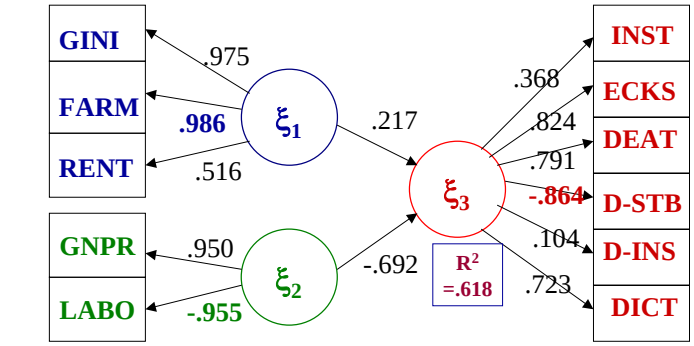
$R^2 = 0.618$

Instabilité politique
= 0.217×Inégalité agricole – 0.692×Développement industriel
(2.24) (-7.22)

(t de Sudent de la régression multiple)

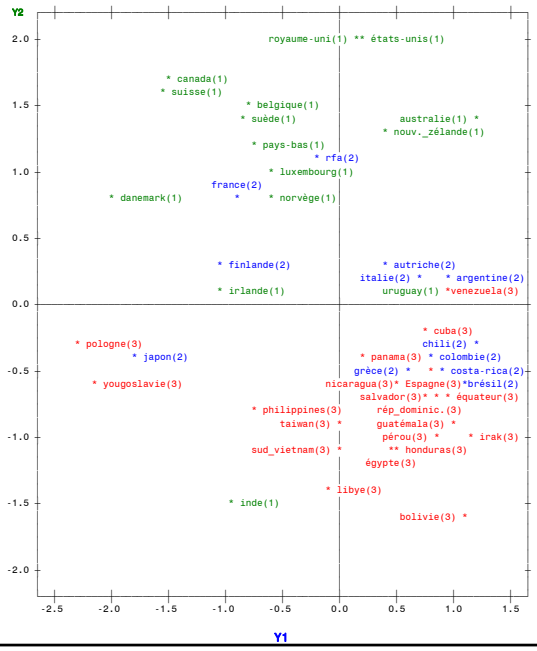
146/200

Inégalité économique et instabilité politique



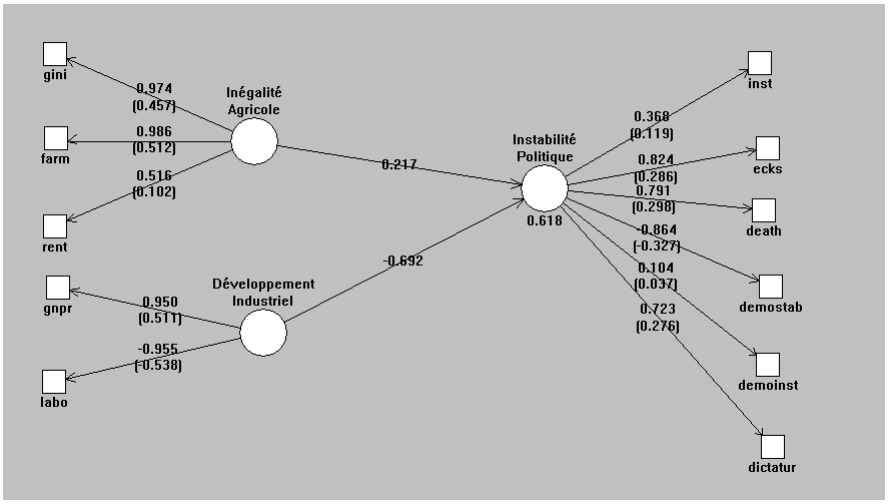
147/200

Carte des pays : Y_1 = inégalité agricole , Y_2 = développement industriel



148/200

Utilisation de PLS-Graph



Utilisation de PLS-Graph

Validation Bootstrap

Outer Model Weights:

	Entire sample estimate	Mean of subsamples	Standard error	T-Statistic
Inégalité agricole:				
gini	0.4567	0.4514	0.0504	9.0674
farm	0.5125	0.5107	0.0519	9.8810
rent	0.1018	0.0862	0.1989	0.5118
Développement industriel:				
gnpr	0.5113	0.5136	0.0246	20.8030
labo	-0.5384	-0.5375	0.0251	-21.4424
Instabilité politique:				
inst	0.1187	0.0992	0.0715	1.6604
ecks	0.2855	0.2765	0.0288	9.9173
demostab	-0.3271	-0.3261	0.0367	-8.9145
demoinst	0.0370	0.0306	0.0595	0.6223
dictatur	0.2758	0.2803	0.0362	7.6143
death	0.2977	0.2940	0.0319	9.3465

Utilisation de PLS-Graph

Validation Bootstrap

Outer Model Loadings:

	Entire sample estimate	Mean of subsamples	Standard error	T-Statistic
Inégalité agricole:				
gini	0.9745	0.9584	0.0336	28.9616
farm	0.9857	0.9689	0.0329	29.9339
rent	0.5156	0.4204	0.2462	2.0946
Développement industriel:				
gnpr	0.9501	0.9489	0.0121	78.3692
labo	-0.9551	-0.9536	0.0107	-89.1493
Instabilité politique:				
inst	0.3676	0.3347	0.1756	2.0932
ecks	0.8241	0.8138	0.0699	11.7920
demostab	-0.8635	-0.8520	0.0667	-12.9419
demoinst	0.1037	0.0955	0.1611	0.6438
dictatur	0.7227	0.7195	0.0841	8.5915
death	0.7910	0.7977	0.0528	14.9773

151/200

PLS-Graph : Validation Bootstap

Path Coefficients Table (Entire Sample Estimate):

	Inég. Agric.	Dev. Indust.	Instab. Pol.
Inég. Agric.	0.0000	0.0000	0.0000
Dev. Indust.	0.0000	0.0000	0.0000
Inst. Pol.	0.2170	-0.6920	0.0000

Path Coefficients Table (Mean of Subsamples):

	Inég. Agric.	Dev. Indust.	Instab. Pol.
Inég. Agric.	0.0000	0.0000	0.0000
Dev. Indust.	0.0000	0.0000	0.0000
Inst. Pol.	0.2328	-0.6743	0.0000

Path Coefficients Table (Standard Error):

	Inég. Agric.	Dev. Indust.	Instab. Pol.
Inég. Agric.	0.0000	0.0000	0.0000
Dev. Indust.	0.0000	0.0000	0.0000
Instabil	0.1272	0.0900	0.0000

Path Coefficients Table (T-Statistic)

	Inég. Agric.	Dev. Indust.	Instab. Pol.
Inég. Agric.	0.0000	0.0000	0.0000
Dev. Indust.	0.0000	0.0000	0.0000
Inst. Pol.	1.7054	-7.6855	0.0000

152/200

Validation de l'uni-dimensionalité d'un bloc (Fiabilité de l'outil de mesure)

1. AVE (Average Variance Explained)

De

$$X_j = \lambda_j \xi + \varepsilon_j$$

et

$$\sum \text{Var}(X_j) = \sum \lambda_j^2 \text{Var}(\xi) + \sum \text{Var}(\varepsilon_j)$$

et Var(ξ) = 1, on déduit :

$$\text{AVE} = \frac{\sum \lambda_j^2}{\sum \text{Var}(X_j)}$$

Règle : AVE > 50%

153/200

Validation de l'uni-dimensionalité d'un bloc

2. Indice de concordance (Composite Reliability)

De

$$X_j = \lambda_j \xi + \varepsilon_j$$

et

$$\sum X_j = \sum \lambda_j \xi + \sum \varepsilon_j$$

et Var(ξ) = 1, on déduit :

$$\text{IC} = \frac{(\sum \lambda_j)^2}{\text{Var}(\sum X_j)} = \frac{(\sum \lambda_j)^2}{(\sum \lambda_j)^2 + \sum \text{Var}(\varepsilon_j)}$$

Pour interpréter cet indice il faut supposer tous les $\lambda_j > 0$

Règle : IC > .70

154/200

Validation de l'uni-dimensionalité d'un bloc

3. Validité convergente

La corrélation entre chaque variable manifeste et sa variable latente doit être supérieure à 0.7 en valeur absolue

155/200

Validité discriminante

1) Une variable manifeste doit être plus corrélée à sa propre variable latente qu'aux autres variables latentes

2) Chaque variable latente doit mieux expliquer ses propres variables manifestes que chaque autre variable latente :

$$AVE(\xi_h) > Cor^2(\xi_h, \xi_k) \text{ pour } k \neq h$$

156/200

LES CAS PARTICULIERS DE LA METHODE PLS

- Analyse en composantes principales
- Analyse factorielle multiple
- Analyse canonique
- Analyse des redondances (ACPVI)
- Régression PLS
- Analyse canonique généralisée (Horst)
- Analyse canonique généralisée (Carroll)

157/200

Les options de l'algorithme PLS

Estimation externe

$Y_j = X_j w_j$

Mode A :

$w_{jh} = cor(X_{jh} , Z_j)$

Mode B :

$w_j = (X_j'X_j)^{-1}X_j'Z_j$

Estimation interne

$$Z_j = \sum e_{ji} Y_i$$

Schéma centroïde

e_{ji} = signe $cor(Y_i,Y_j)$

Schéma factoriel

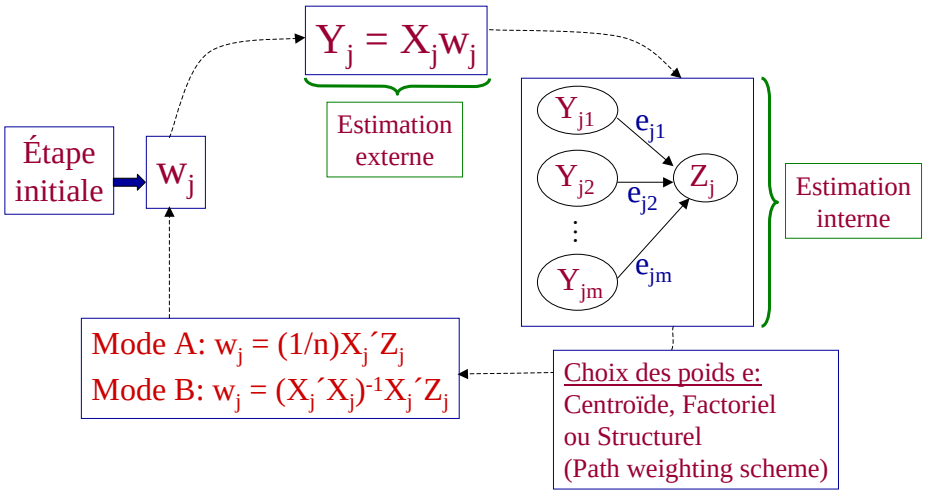
$e_{ji} = cor(Y_i,Y_j)$

Schéma structurel

e_{ji} = coeff. de régression dans la régression de Y_j sur les Y_i

158/200

L'algorithme PLS général



159/200

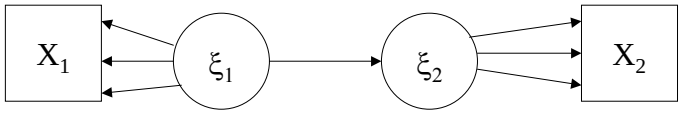
Modification de méthodes d'analyse multi-bloc pour l'analyse des modèles de causalité

c_{jk} = 1 si les blocs sont reliés, = 0 sinon

SUMCOR (Horst, 1961)	$Max \sum_{j,k} c_{jk} Cor(Y_j, Y_k)$
Mathes (1993), Hanafi (2004):	$Max \sum_{j,k} c_{jk} Cor^2(Y_j, Y_k)$
Mathes (1993), Hanafi (2004)	$Max \sum_{j,k} c_{jk} \lvert Cor(Y_j, Y_k) \rvert$
MAXBET (Van de Geer, 1984 & Ten Berge, 1988):	$\underset{All \lVert w_j \rVert=1}{Max} [\sum_j Var(X_j w_j) + \sum_{j \neq k} c_{jk} Cov(X_j w_j, X_k w_k)]$
MAXDIFF (Van de Geer, 1984 & Ten Berge, 1988):	$\underset{All \lVert w_j \rVert=1}{Max} [\sum_{j \neq k} c_{jk} Cov(X_j w_j, X_k w_k)]$

Mathes-Hanafi 1 ⇒ Mode B and Schéma factoriel
Mathes-Hanafi 2 ⇒ Mode B and Schéma centroïde

Approche PLS : 2 blocs

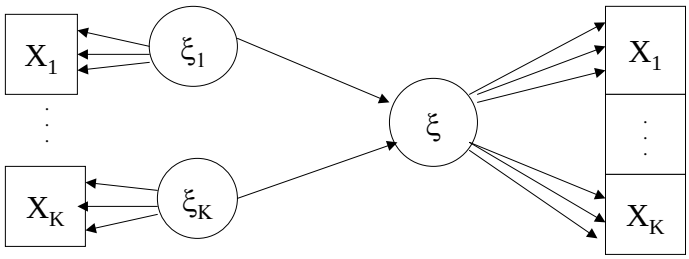


Mode de calcul des w_j

$Y_1 = X_1w_1$	$Y_2 = X_2w_2$	Méthode	Déflation
A	A	Régression PLS de X_2 sur X_1	Sur X_1 seulement
B	A	Analyse des redondances de X_2 par rapport à X_1	Sur X_1 seulement
A	A	Analyse Factorielle Inter-Batteries de Tuker	Sur X_1 et X_2
B	B	Analyse canonique	Sur X_1 et X_2

161/200

Approche PLS : K blocs



Mode de calcul de $Z_j = \sum e_{ji} Y_i$

Mode de calcul des w_i	**Centroïde**	**Factoriel**	**Structurel**
A	**ACG de Horst PLS**	**ACG de Carroll PLS**	- **ACP de X** - **AFM des X_i**
B	**ACG de Horst (SUMCOR)**	**ACG de Carroll**	Nouveau

Déflation :
sur le super-bloc seulement

162/200

Approche PLS

Exemple d'utilisation de l'approche PLS pour l'analyse de tableaux multiples

(Christiane Guinot et Michel Tenenhaus)

Étude des habitudes de consommation de produits cosmétiques des femmes d'Ile de France

163/200

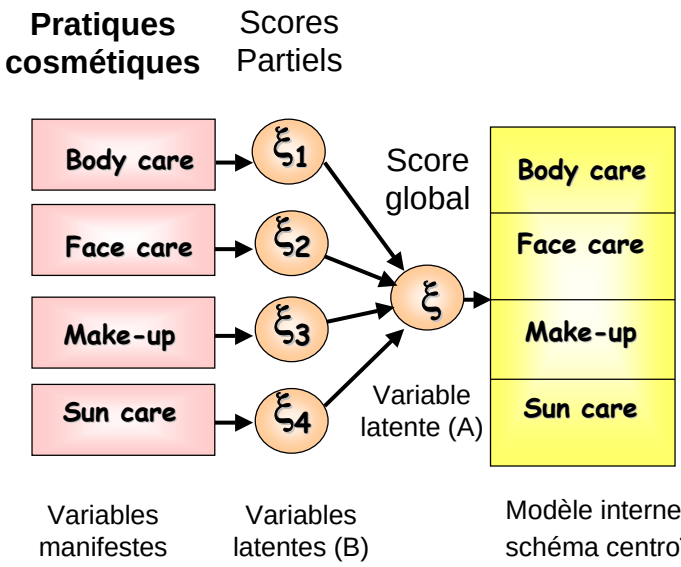
Les données

Les produits cosmétiques ont été divisés en quatre blocs correspondant à différentes habitudes d'utilisation de produits cosmétiques.

Body care	soap, liquid soap, moisturising body cream, hand creams
Face care	make-up and eye make-up removers, tonic lotions, day creams, night creams exfoliation products
Make -up	blushers, mascaras, eye shadows, eye pencils, lipsticks, lip shiners and nail polish
Sun care	sun protection products for face and for body after-sun products for face and for body

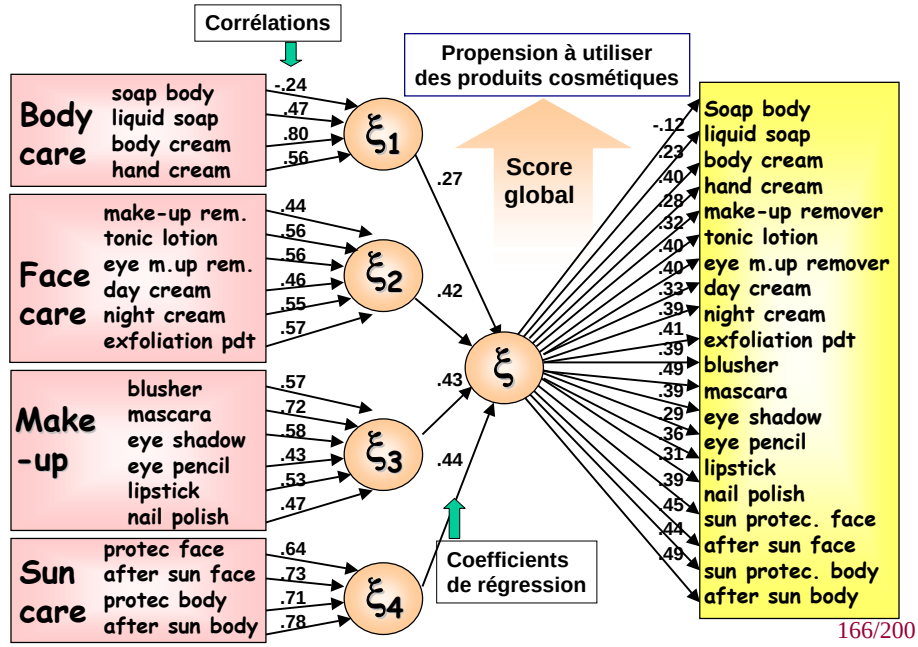
164/200

Construction d'un score global



165/200

Résultats



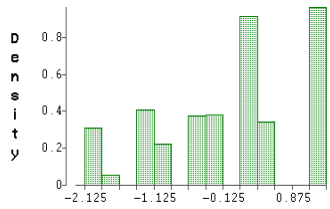
166/200

Calcul du score global

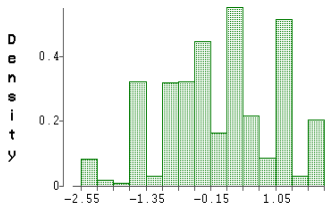
Score global = -3.40
- .11 * soaps and toilet soaps for body care
+.20 * liquid soaps for body care
+.38 * moisturising body creams and milks
+.25 * hand creams and milks
+.21 * make-up removers
+.26 * tonic lotions
+.30 * eye make-up removers
+.39 * moisturising day creams
+.30 * moisturising night creams
+.30 * exfoliation products
+.26 * blushers
+.41 * mascaras
+.26 * eye shadows
+.20 * eye pencils
+.33 * lipsticks and lip shiners
+.20 * nail polish
+.36 * sun protection products for the face
+.31 * moisturising after sun products for the face
+.38 * sun protection products for the body
+.34 * moisturising after sun products for the body

167/200

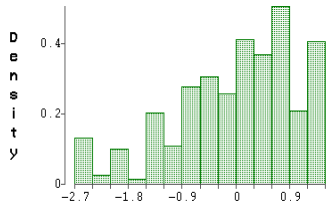
Histogrammes des scores partiels



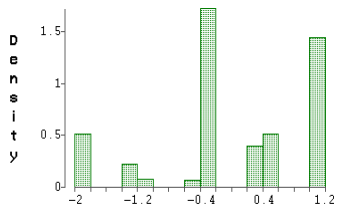
Score body-care



Score facial-care



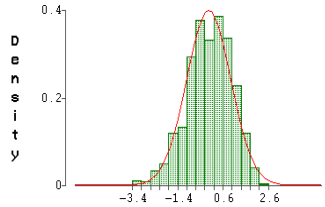
Score make-up



Score sun-care

168/200

Histogramme du score global



	S_body-care	S_facial-care	S_make-up	S_sun-care
S_facial-care	0.24001			
S_make-up	0.13462	0.35035		
S_sun-care	0.16500	0.19075	0.14273	
S_global	0.50263	0.71846	0.67347	0.62071

169/200

Facteurs influençant l'utilisation de produits cosmétiques

On peut relier le score global d'utilisation de produits cosmétiques à des caractéristiques décrivant les consommatrices :

- Activité professionnelle et CSP
- Enfants
- Habitudes d 'exposition solaire
- Pratiques sportives
- Importance de l 'apparence physique
- Type de peau (visage et corps)
- Age

170/200

Score global en fonction des caractéristiques des consommatrices

E(Score global) = -1.02
+.21 * professional activity
+.07 * housewife or student
+.00 * retired
+.27 * CSP A (craftsmen, trades people, business managers, managerial staff,academics and professionals)
+.09 * CSP B (farmers and intermediary professions)
+.05 * CSP C (employees and working class people)
+.00 * CSP D (retired and non working people)
-.21 * without child
+.00 * with child
+.40 * habits of deliberate exposure to sunlight
+.09 * previous habits of deliberate exposure to sunlight
+.00 * no habits of deliberate exposure to sunlight
-.17 * no sport practised
+.00 * sport practised
+1.04 * physical appearance is of extreme importance
+.89 * physical appearance is of high importance
+.50 * physical appearance is of some importance
+.00 * physical appearance is of little importance
-.06 * oily facial skin
+.16 * combination facial skin
-.20 * normal facial skin
+.00 * dry facial skin
-.32 * oily body skin
-.57 * combination body skin
-.32 * normal body skin
+.00 * dry body skin
-.00 * age

171/200

Exemple d'un bon profil

1.06

E(Global score)= -1.02
+.21 * professional activity
+.07 * housewife or student
+.00 * retired
+.27 * CSP A (craftsmen, trades people, business managers, managerial staff, academics and professionals)
+.09 * CSP B (farmers and intermediary professions)
+.05 * CSP C (employees and working class people)
+.00 * CSP D (retired and non working people)
- .21 * without child
+.00 * with child
+.40 * habits of deliberate exposure to sunlight
+.09 * previous habits of deliberate exposure to sunlight
+.00 * no habits of deliberate exposure to sunlight
- .17 * no sport practised
+.00 * sport practised
+1.04 * physical appearance is of extreme importance
+.89 * physical appearance is of high importance
+.50 * physical appearance is of some importance
+.00 * physical appearance is of little importance
- .06 * oily facial skin
+.16 * combination facial skin
- .20 * normal facial skin
+.00 * dry facial skin
- .32 * oily body skin
- .57 * combination body skin
- .32 * normal body skin
+.00 * dry body skin
- .00 * age

172/200

Exemple d'un profil non cible

E(Score global)= -1.02

-2.00

+.21 * professional activity
+.07 * housewife or student
+.00 * retired
+.27 * CSP A (craftsmen, trades people, business managers, managerial staff, academics and professionals)
+.09 * CSP B (farmers and intermediary professions)
+.05 * CSP C (employees and working class people)
+.00 * CSP D (retired and non working people)
-.21 * without child
+.00 * with child
+.40 * habits of deliberate exposure to sunlight
+.09 * previous habits of deliberate exposure to sunlight
+.00 * no habits of deliberate exposure to sunlight
-.17 * no sport practised
+.00 * sport practised
+1.04 * physical appearance is of extreme importance
+.89 * physical appearance is of high importance
+.50 * physical appearance is of some importance
+.00 * physical appearance is of little importance
-.06 * oily facial skin
+.16 * combination facial skin
-.20 * normal facial skin
+.00 * dry facial skin
-.32 * oily body skin
-.57 * combination body skin
-.32 * normal body skin
+.00 * dry body skin
-.00 * age

Conclusion

L'utilisation de l'approche PLS a permis d'obtenir un score de la propension à utiliser des produits cosmétiques en équilibrant les différents types de produits cosmétiques de manière plus efficace que l 'analyse en composantes principales.

Les méthodes PLS

VI. Régression logistique PLS

175/200

Qualité des vins de Bordeaux

Variables observées sur 34 années (1924 - 1957)

- TEMPERATURE : Somme des températures moyennes journalières
- SOLEIL : Durée d'insolation
- CHALEUR : Nombre de jours de grande chaleur
- PLUIE : Hauteur des pluies
- QUALITE DU VIN : Bon, Moyen, Médiocre

176/200

Les données

	Température	Soleil	Chaleur	Pluie	Qualité
1	3064	1201	10	361	2
2	3000	1053	11	338	3
3	3155	1133	19	393	2
4	3085	970	4	467	3
5	3245	1258	36	294	1
6	3267	1386	35	225	1
7	3080	966	13	417	3
8	2974	1189	12	488	3
9	3038	1103	14	677	3
10	3318	1310	29	427	2
11	3317	1362	25	326	1
12	3182	1171	28	326	3
13	2998	1102	9	349	3
14	3221	1424	21	382	1
15	3019	1230	16	275	2
16	3022	1285	9	303	2
17	3094	1329	11	339	2
18	3009	1210	15	536	3
19	3227	1331	21	414	2
20	3308	1366	24	282	1
21	3212	1289	17	302	2
22	3361	1444	25	253	1
23	3061	1175	12	261	2
24	3478	1317	42	259	1
25	3126	1248	11	315	2
26	3458	1508	43	286	1
27	3252	1361	26	346	2
28	3052	1186	14	443	3
29	3270	1399	24	306	1
30	3198	1259	20	367	1
31	2904	1164	6	311	3
32	3247	1277	19	375	1
33	3083	1195	5	441	3
34	3043	1208	14	371	3

177/200

Régression logistique ordinale

Y = Qualité : Bon (1), Moyen (2), Médiocre (3)

PROB(Y ≤ i) =

$$\frac{e^{\alpha_i+\beta_1\text{Température}+\beta_2\text{Soleil}+\beta_3\text{Chaleur}+\beta_4\text{Pluie}}}{1+e^{\alpha_i+\beta_1\text{Température}+\beta_2\text{Soleil}+\beta_3\text{Chaleur}+\beta_4\text{Pluie}}}$$

178/200

Régression logistique ordinale
Résultats SAS

Score Test for the Proportional Odds Assumption

Chi-Square = 2.9159 with 4 DF (p=0.5720)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCP1	1	-2.6638	0.9266	8.2641	0.0040
INTERCP2	1	2.2941	0.9782	5.4998	0.0190
TEMPERA	1	3.4268	1.8029	3.6125	0.0573
SOLEIL	1	1.7462	1.0760	2.6335	0.1046
CHALEUR	1	-0.8891	1.1949	0.5536	0.4568
PLUIE	1	-2.3668	1.1292	4.3931	0.0361

179/200

Régression logistique ordinale
Qualité de prévision du modèle

QUALITE OBSERVEE Effectif	PREVISION 1	2	3	Total
1	8	3	0	11
2	2	8	1	11
3	0	1	11	12
Total	10	12	12	34

Résultat : 7 années mal classées

180/200

Régression logistique ordinale
Commentaires

- Le modèle à pentes égales est acceptable (p = 0.572).
- La chaleur a une influence positive sur la qualité du vin de Bordeaux, alors qu'elle apparaît comme non significative et avec un coefficient négatif dans le modèle.
- C 'est un problème de multicolinéarité.
- Il y a 7 années mal classées.

181/200

Régression logistique PLS

- Bonne solution au problème de la multicolinéarité.
- Il peut y avoir beaucoup plus de variables que d'observations.
- Il peut y avoir des données manquantes.
- Présentation de deux algorithmes.

182/200

Algorithme 1 :
La régression logistique PLS avec sélection de variables

Etape 1 : Recherche de m composantes orthogonales $T_h = Xa_h$ explicatives de leur propre groupe et bien prédictives de y.

Le nombre m correspond au nombre de composantes significatives.

Etape 2 : Régression logistique de Y sur les composantes T_h .

Etape 3 : Expression de la régression logistique en fonction de X.

183/200

Régression logistique PLS
Étape 1 : Construction de T_1

1. Régression logistique de y sur chaque x_j :

   $\Rightarrow$ les coefficients de régression a_{1j}

   Les coefficients de régression a_{1j} non significatifs sont mis à 0. Seules les variables significatives contribuent à la construction de T_1.

2. Normalisation du vecteur $a_1 = (a_{11},\ldots,a_{1k})$
3. Calcul de $T_1 = Xa_1$
4. Régression logistique de y sur $T_1 = Xa_1$ exprimée en fonction des X

184/200

Application Bordeaux
Étape 1 : Construction de T_1

Les quatre régressions logistiques :

	Coefficient	p-value
Température	3.0117	.0002
Soleil	3.3401	.0002
Chaleur	2.1445	.0004
Pluie	-1.7906	.0016

La composante PLS T_1 :

$$T_1 = \frac{3.0117\,\text{Température} + 3.3401\,\text{Soleil} + 2.1445\,\text{Chaleur} - 1.7906\,\text{Pluie}}{\sqrt{(3.0117)^2 + (3.3401)^2 + (2.1445)^2 + (-1.7906)^2}}$$

$$= 0.5688\,\text{Température} + 0.6309\,\text{Soleil} + 0.4050\,\text{Chaleur} - 0.3382\,\text{Pluie}$$

185/200

Application Bordeaux
Étape 2 : Régression logistique sur T_1

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-2.2650	0.8644	6.8662	0.0088
Intercept2	1	2.2991	0.8480	7.3497	0.0067
t1	1	2.6900	0.7155	14.1336	0.0002

TABLEAU CROISANT QUALITÉ OBSERVÉE ET PRÉDITE

QUALITÉ PRÉDICTION

Effectif	1	2	3	Total
1	9	2	0	11
2	2	8	1	11
3	0	1	11	12
Total	11	11	12	34

6 mal classés

186/200

Application Bordeaux
Étape 3 : Régression logistique en fonction des X

$$\text{Prob}(Y = 1) = \frac{e^{-2.265+1.53\times\text{Température}+1.70\times\text{Soleil}+1.09\times\text{Chaleur}-.91\times\text{Pluie}}}{1+e^{-2.5265+1.53\times\text{Température}+1.70\times\text{Soleil}+1.09\times\text{Chaleur}-.91\times\text{Pluie}}}$$

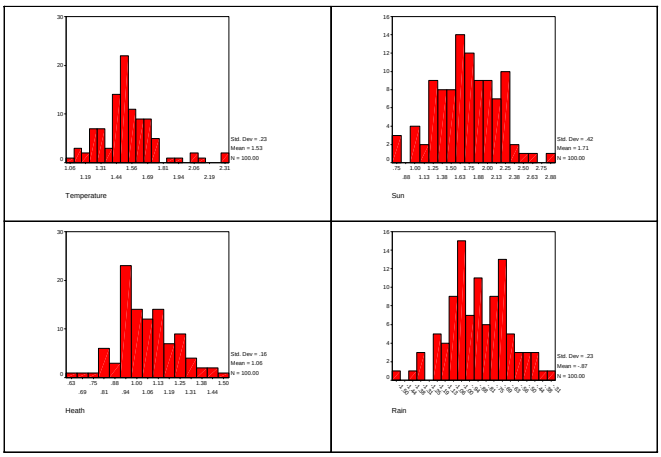
et

$$\text{Prob}(Y \le 2) = \frac{e^{2.2991+1.53\times\text{Température}+1.70\times\text{Soleil}+1.09\times\text{Chaleur}-.91\times\text{Pluie}}}{1+e^{2.2991+1.53\times\text{Température}+1.70\times\text{Soleil}+1.09\times\text{Chaleur}-.91\times\text{Pluie}}}$$

Commentaires : Ce modèle est plus cohérent au niveau des coefficients de régression que le modèle de régression logistique ordinale usuelle et conduit ici à un mal classé de moins sur l'échantillon utilisé.

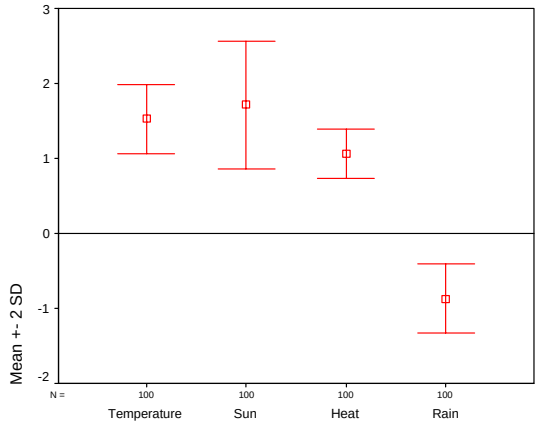
187/200

Application Bordeaux
Validation Bootstrap du modèle à une composante (100 échantillons)



188/200

Application Bordeaux
Validation Bootstrap du modèle à une composante
Intervalle de confiance à 95% des coefficients



189/200

Régression logistique PLS
Construction de T_2

1. Régression logistique de y sur T_1 et chaque variable x_j.
   Pour construire T_2 on ne sélectionne que les variables x_j significatives.
2. On construit les résidus x_{1j} des régressions des x_j sélectionnés sur T_1.
3. On construit les régressions logistiques de y sur T_1 et chaque x_{1j} retenu.
   $\Rightarrow$ les coefficients de régression b_{2j} de x_{1j}.
4. Normalisation du vecteur $b_2 = (b_{21},\ldots,b_{2k})$
5. Calcul de a_2 tel que : $T_2 = X_1 b_2 = X a_2$
6. Régression logistique de y sur $T_1 = X a_1$ et $T_2 = X a_2$ exprimée en fonction des X

190/200

Application Bordeaux
Sélection des variables contribuant à la construction de T_2

Régression logistique de la qualité sur T_1 et chaque X

	Coefficient	p-value
Température	-.6309	.6765
Soleil	.6459	.6027
Chaleur	-1.9407	.0983
Pluie	-.9798	.2544

Commentaires : En choisissant un risque de 5% on décide donc de ne conserver qu'une seule composante PLS.

191/200

Algorithme 2
Régression logistique sur composantes PLS

(1) Régression PLS des indicatrices de Y sur les X.

(2) Régression logistique de Y sur les composantes PLS des X.

192/200

Régression logistique sur les composantes PLS
Résultats

- La température de 1924 est supposée inconnue.
- La régression PLS des indicatrices de Y sur X a conduit à une seule composante PLS t_1 (résultat de la validation croisée).
- $t_1 = 0.55\times Température + 0.55\times Soleil + 0.48\times Chaleur - 0.40\times Pluie$
- Pour l'année 1924 :

$$t_1 = (0.55\times Soleil + 0.48\times Chaleur - 0.40\times Pluie)/0.69$$

Utilisation de la régression PLS pour la prévision de la qualité du vin de Bordeaux

The PLS Procedure
Cross Validation for the Number of Latent Variables

Number of Latent Variables	Root Mean PRESS	Prob > PRESS
0	1.0313	0
1	0.8304	1.0000
2	0.8313	0.4990
3	0.8375	0.4450
4	0.8472	0.3500

Minimum Root Mean PRESS = 0.830422 for 1 latent variable
Smallest model with p-value > 0.1: 1 latent

TABLE OF QUALITE BY PREV

QUALITE / Frequency	1	3	Total
1	11	0	11
2	4	7	11
3	1	11	12
Total	16	18	34

Choix d'une composante PLS

Résultat :
12 années mal classées

Résultats de la régression logistique de Y sur la composante PLS t_1

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCP1	1	-2.1492	0.8279	6.7391	0.0094
INTERCP2	1	2.2845	0.8351	7.4841	0.0062
t1	1	2.6592	0.7028	14.3182	0.0002

TABLEAU CROISANT QUALITÉ OBSERVÉE ET PRÉDITE

QUALITÉ PRÉDICTION

Effectif	1	2	3	Total
1	9	2	0	11
2	2	8	1	11
3	0	1	11	12
Total	11	11	12	34

Résultat :
6 années mal classées

195/200

Régression logistique sur composantes PLS
Le modèle

Prob ($Y \leq i$)

$$= \frac{e^{-2.15\times Bon+2.28\times Moyen+2.66\times t_1}}{1+e^{-2.15\times Bon+2.28\times Moyen+2.66\times t_1}}$$

$$= \frac{e^{-2.15\times \mathrm{Bon}+2.28\times \mathrm{Moyen}+1.47\times \mathrm{Temp.}+1.46\times \mathrm{Soleil}+1.28\times \mathrm{Chaleur}-1.07\times \mathrm{Pluie}}}{1+e^{-2.15\times \mathrm{Bon}+2.28\times \mathrm{Moyen}+1.47\times \mathrm{Temp.}+1.46\times \mathrm{Soleil}+1.28\times \mathrm{Chaleur}-1.07\times \mathrm{Pluie}}}$$

196/200

Conclusion 1: Régression logistique PLS vs régression logistique sur composantes PLS

- Les deux algorithmes présentés devraient avoir des qualités comparables.
- L ’algorithme 2 est beaucoup plus simple :

  Deux étapes :

  (1) Régression PLS des indicatrices de Y sur X

  (2) Régression logistique de Y sur les composantes PLS

197/200

Conclusion 2: La régression linéaire généralisée PLS

- La régression linéaire généralisée PLS peut être construit selon les mêmes procédures.
- Approche beaucoup plus simple que la méthode de Brian Marx : « Iteratively Reweighted Partial Least Square Estimation for Generalized Linear Regression », *Technometrics*, 1996.

198/200

Quelques références sur les méthodes PLS

Régression PLS

- L. Eriksson, E. Johansson, N. Kettaneh-Wold & S. Wold : *Multi- and Megavariate Data Analysis using Projection Methods (PCA & PLS), 2nd edition* Umetrics, 2005.
- H. Martens & M. Martens : *Multivariate Analysis of Quality*, Wiley, 2000
- SIMCA-P 10.0 : PLS Software, S. WOLD, UMETRI (Sweden), distribué par SIGMA PLUS, 29 rue Lauriston, 75016 Paris
- M. Tenenhaus : *La régression PLS*, Editions Technip, 1998
- P. Bastien, V. Esposito Vinzi, M. Tenenhaus : PLS generalized linear regression, *Computational Statistics & Data Analysis*, 2005

Approche PLS (PLS Path modeling)

- J.-B. Lohmöller : *Latent variable path modeling with partial least squares*, Physica-Verlag, 1989
- CHIN W.W. (2003) : *PLS-Graph User's Guide*, C.T. Bauer College of Business, University of Houston, Houston.
- M. Tenenhaus : L'approche PLS, *R.S.A.*, 47 (2), 5-40, 1999
- M. Tenenhaus, V. Esposito Vinzi, Y.-M. Chatelin, C. Lauro : PLS Path modeling, *Computational Statistics & Data Analysis*, 2005

Conclusion générale

The proof of the pudding is in the eating.