Les Méthodes PLS

Pierre-Louis Gonzalez

Michel Tenenhaus

Les méthodes PLS
initiées par Herman et Svante Wold

I. NIPALS (Nonlinear Iterative Partial Least Squares)

II. Régression PLS (Partial Least Squares Regression)

- II.1 PLS1
- II.2 PLS2

III Analyse discriminante PLS

IV. Régression logistique PLS

I. 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.

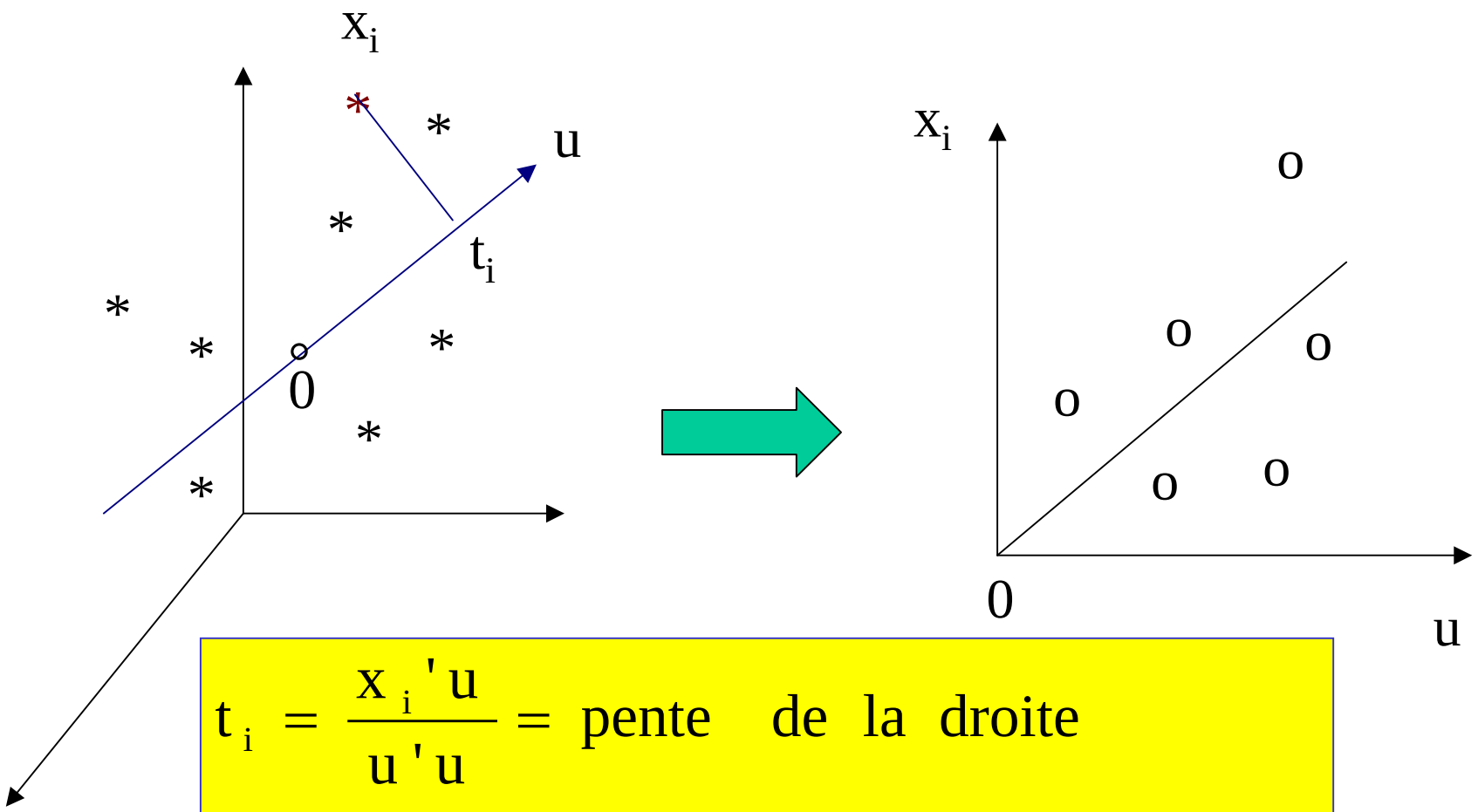
Utilisation de NIPALS :Exemple voitures

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 véhicule !

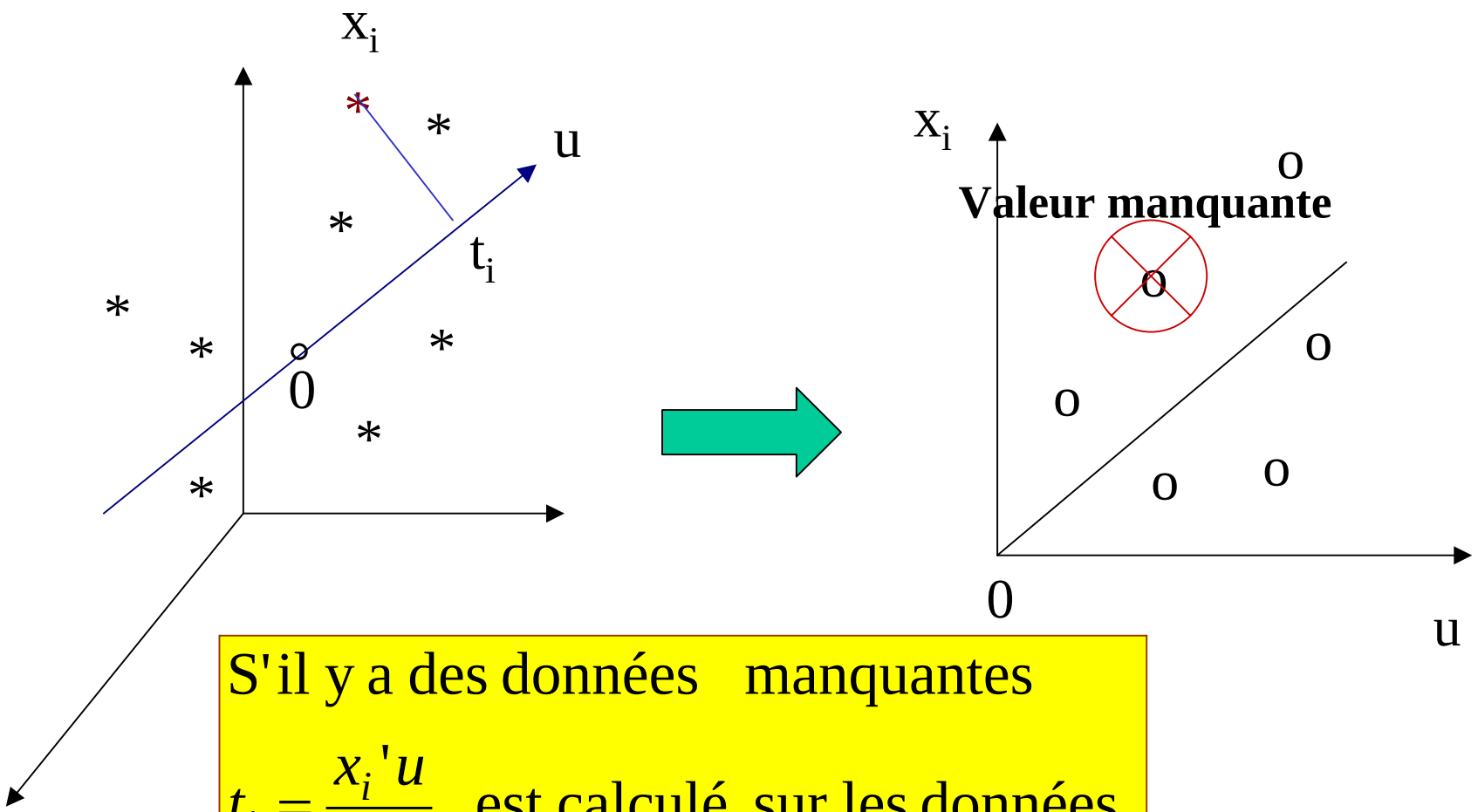
Le principe de NIPALS: Comment projeter un point avec données manquantes ?

Projection sur un axe



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

Projection d'un point avec données manquantes sur un axe



S'il y a des données manquantes

$t_i = \dfrac{x_i'u}{u'u}$ est calculé sur les données disponibles

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 (axes)

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

(composantes principales)

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 :

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

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

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

  $$x_i = t_{1i} p_1 + \text{résidu}$$

- 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.

Commentaires:

Les relations cycliques découlant des équations de base de l'algorithme montrent que λ_1 est la plus grande valeur propre vérifiant les équations suivantes:

$$\frac{1}{n-1} X'X p_1 = \lambda_1 p_1$$

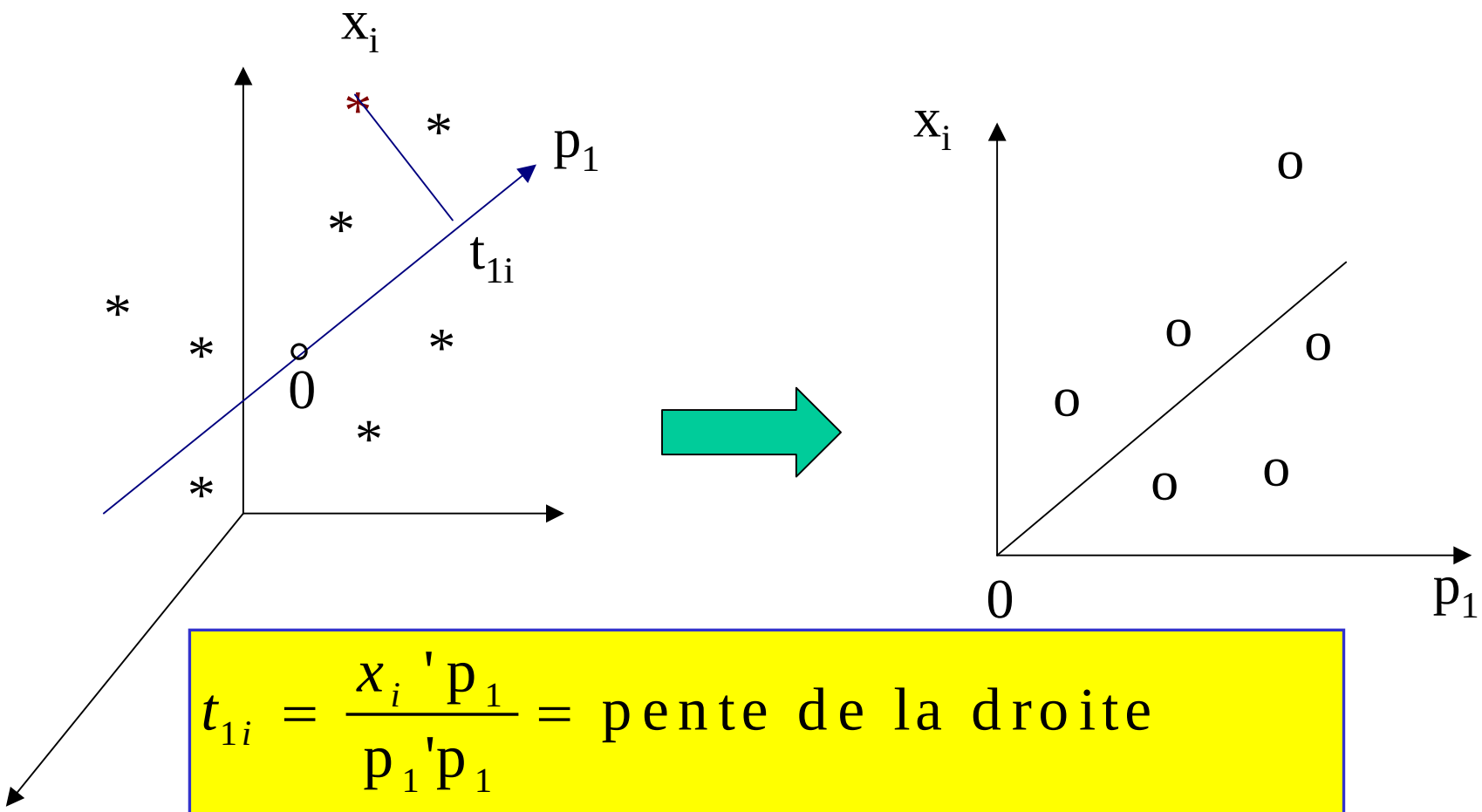
$$\frac{1}{n-1} X X' t_1 = \lambda_1 t_1$$

Nous avons divisé par n-1 pour retrouver les résultats de SIMCA.

Ce calcul est une application de la méthode de la puissance itérée pour le calcul du vecteur propre d'une matrice associé à la plus grande valeur propre

(Hotelling-1936; Anderson-1958)

Projection sur l' axe 1



$$t_{1i} = \frac{x_i' p_1}{p_1' p_1} = \text{pente de la droite}$$

des moindres carrés sans constante

de x_i sur p_1

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.

$RESS_h$ 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.

Predicted Residual Sum of Squares :

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

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

L ’algorithme NIPALS
Choix du nombre de composantes

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

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

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

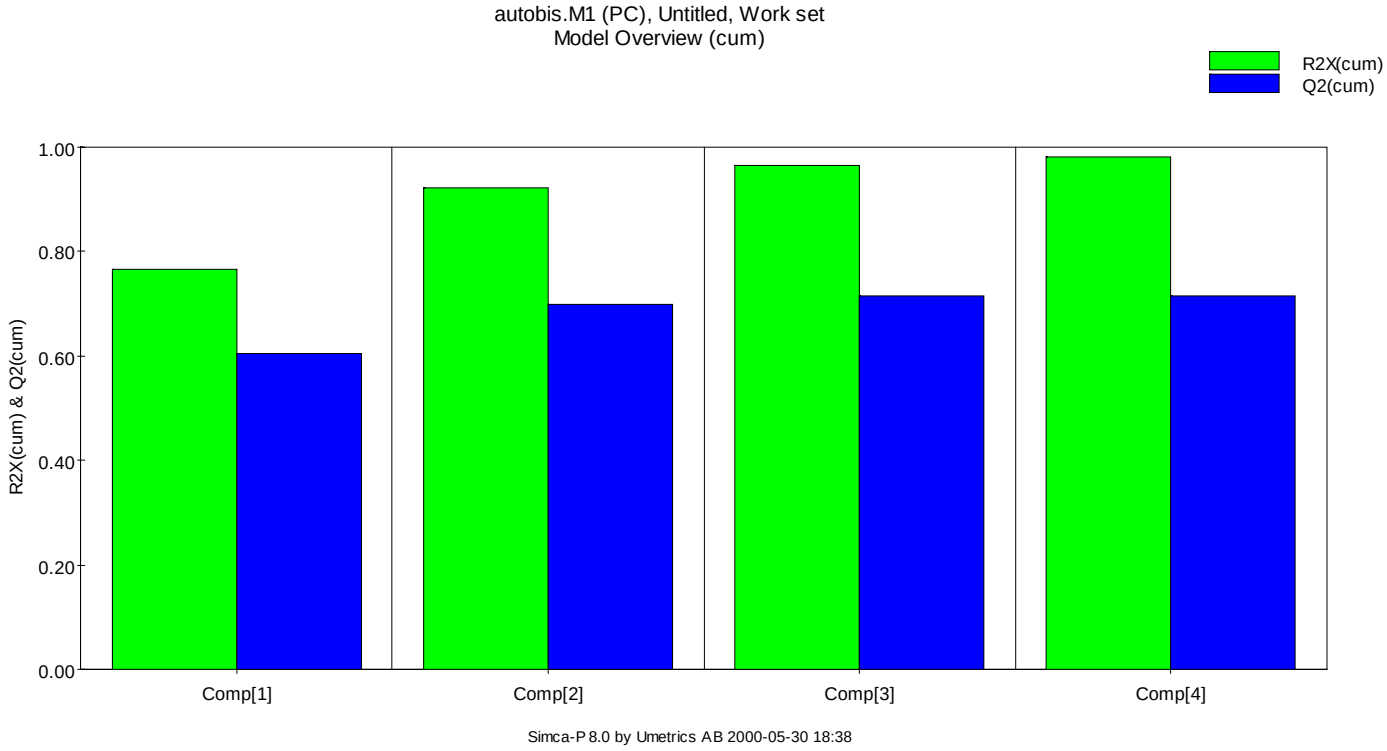
$$R^2_{\text{validation croisée}} = 1 - \frac{PRESS_h / n - 1}{\sum_j s_j^2}$$

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 $[Q^2_{cum}]_h > 0.5$

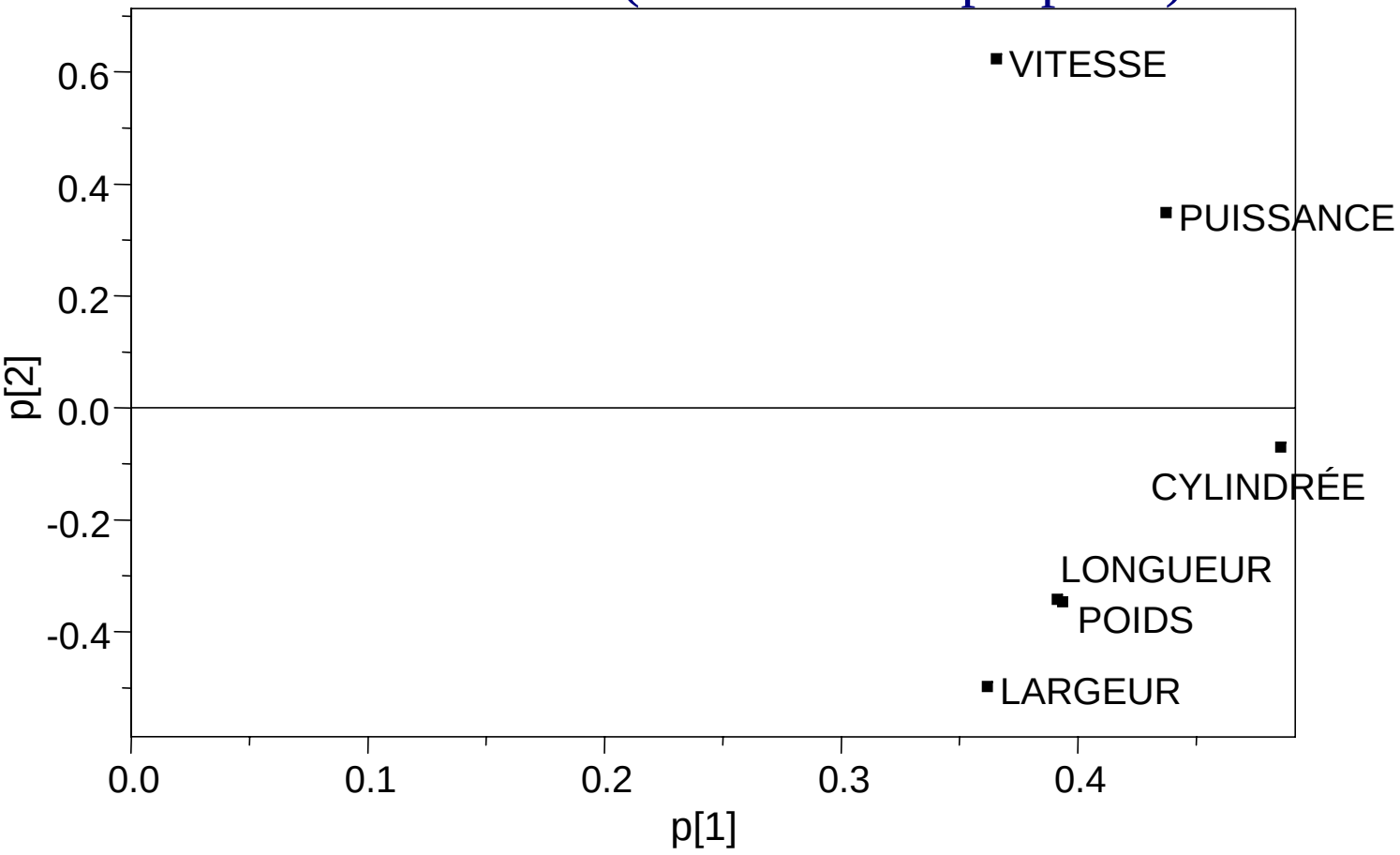
Utilisation de NIPALS : Exemple voitures



La validation croisée conduit à deux composantes.

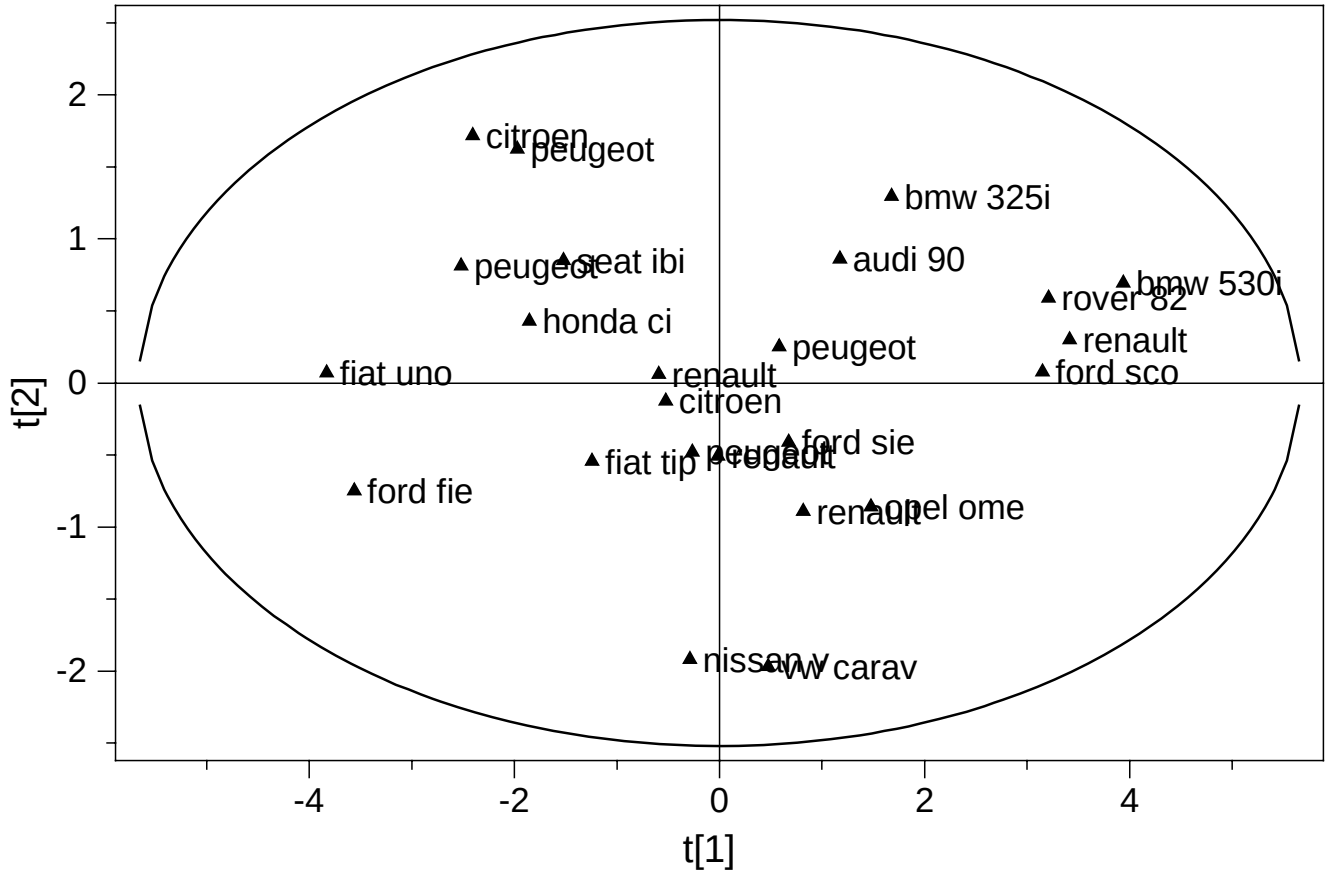
NIPALS : Exemple Voitures

Carte des variables ("les vecteurs propres")



NIPALS : Exemple Voitures

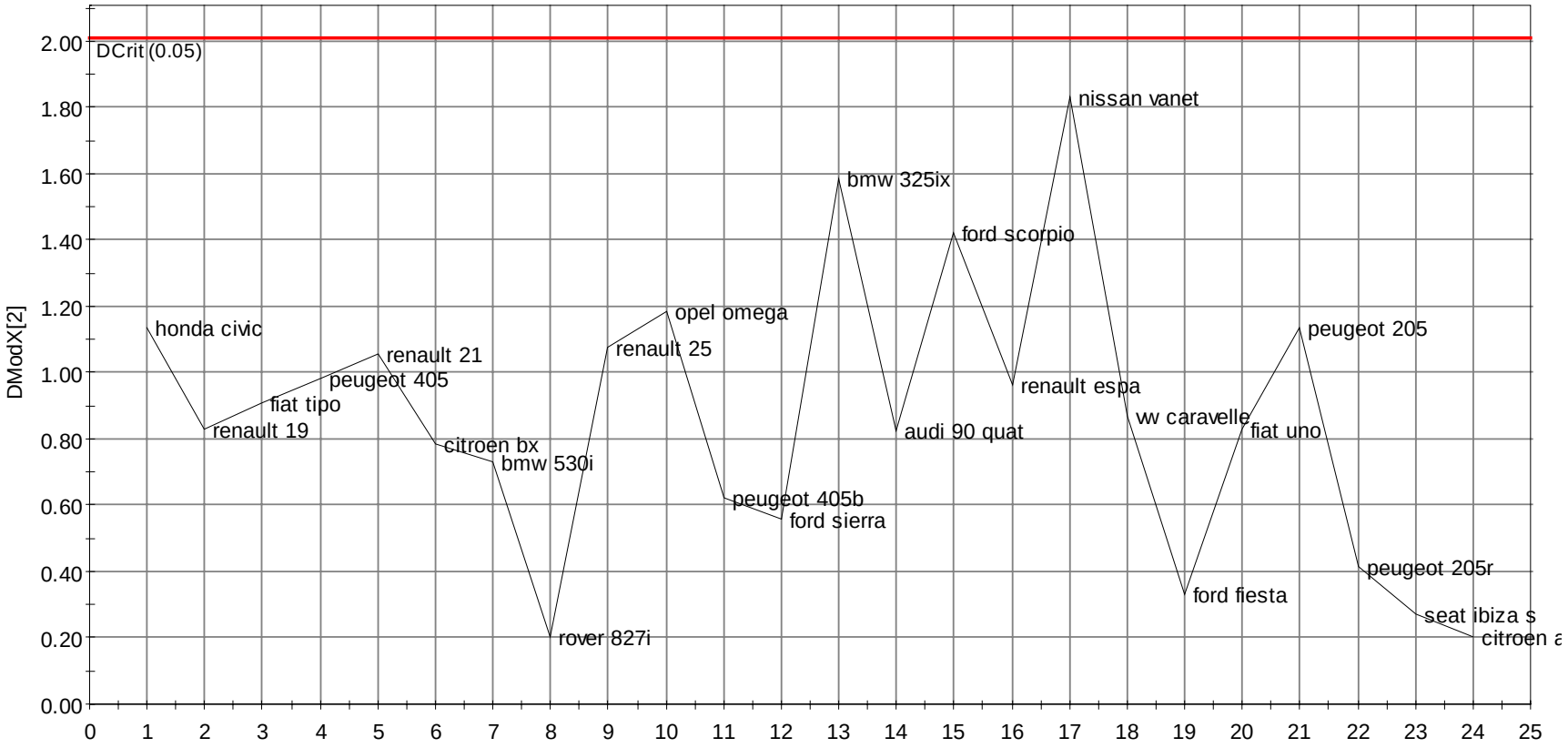
Carte des voitures (les 2 premières "composantes principales")



Ellipse: Hotelling T2 (0.05)

NIPALS : Identification des outliers

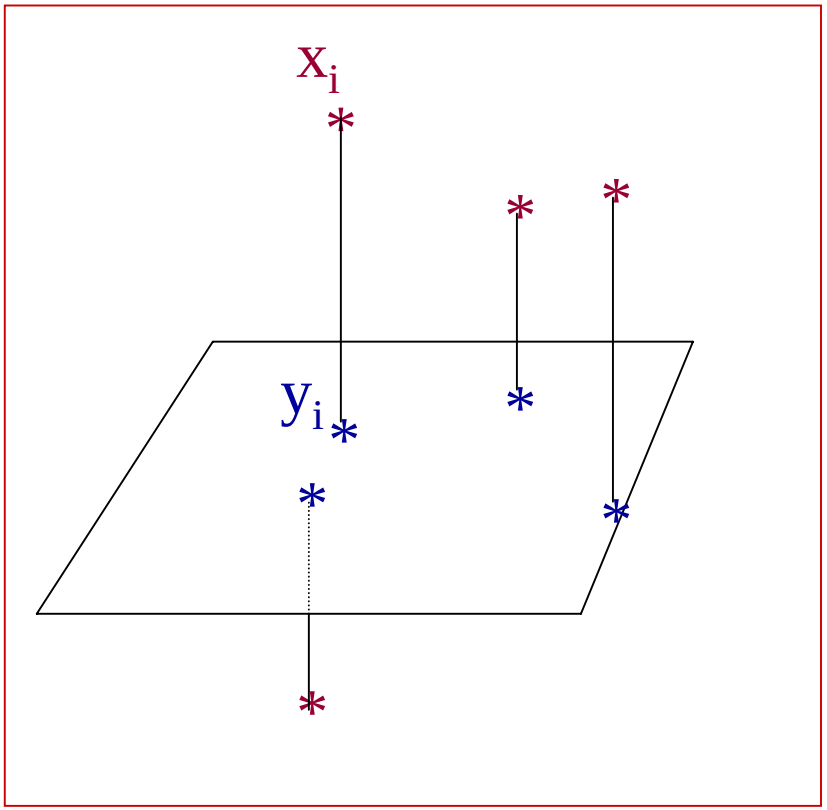
Carte de contrôle des distances au modèle normalisées



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

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

Calcul de la limite de contrôle



Propriété :

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

Limite de contrôle :

$$\sqrt{F_{0.95}(k_1, k_2)}$$

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

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

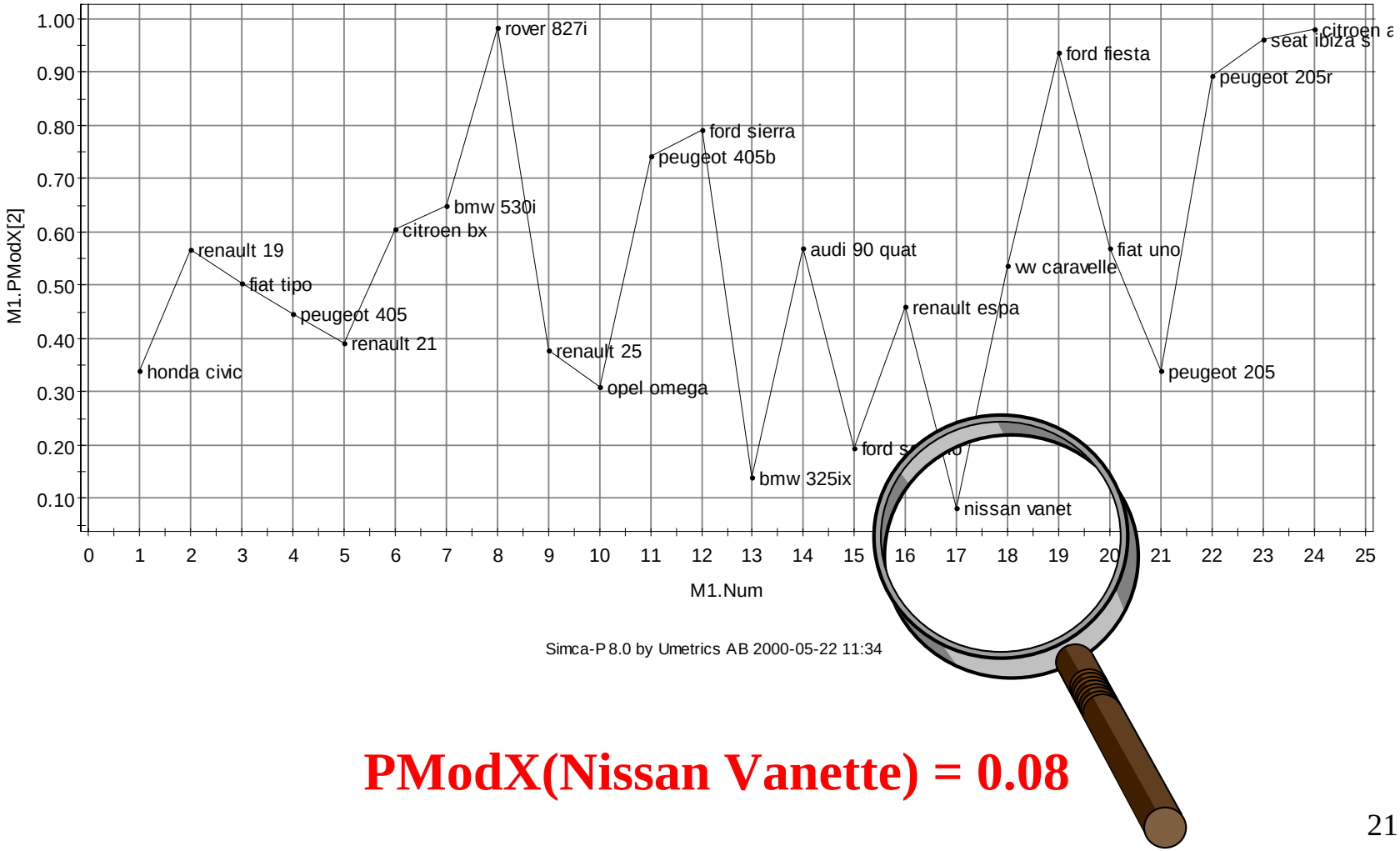
Niveau de signification ou « probabilité d'appartenir au modèle » :

Plus petit α conduisant au rejet de H_0

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

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

NIPALS : Exemple Voitures
"Probabilité" d'appartenir au modèle ACP (2 composantes)



PModX(Nissan Vanette) = 0.08

II. 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.**

La régression PLS : vocabulaire

- **Régression PLS1** : un seul Y
- **Régression PLS2** : plusieurs Y
- **Analyse discriminante PLS** :

  Y qualitatif transformé en variables indicatrices des modalités

II.1. La régression PLS1 :une idée de l'algorithme

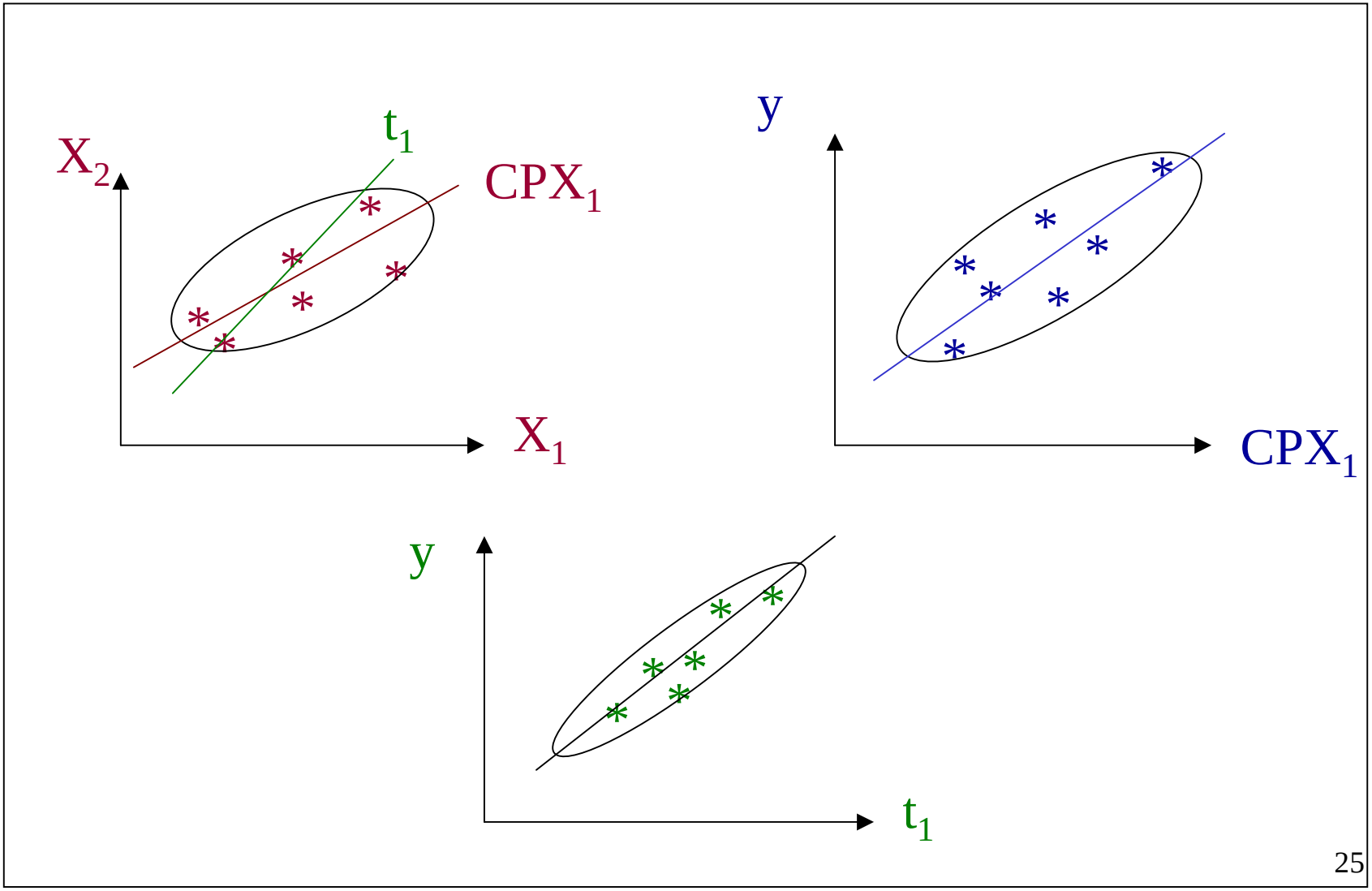
Etape 1 : Recherche de m composantes orthogonales $t_h = Xa_h$ bien explicatives de leur propre groupe et bien corrélées à y.

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.

Objectif de l'étape 1 de la régression PLS1



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 ,…, t_{h-1}.

Propriétés de la régression PLS1

De $Cov^2(Xa_h , y) = Cor^2(Xa_h , y)*Var(Xa_h)*Var(y)$

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.

Régression PLS1: Étape 1

1. Calcul de la première composante PLS t_1 :

$$t_1 = Xa_1 = \sum_j cor(y, x_j) \times x_j$$

Lors de cette étape les covariances sont égales aux corrélations, puisque toutes les données sont centrées réduites

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 :

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

2. Normalisation du vecteur $b_2 = (b_{21},\ldots,b_{2k})$
3. Calcul de a_2 tel que : $t_2 = X_1 b_2 = X a_2$
4. Régression de y_1 sur $t_2 = X a_2$ 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 :

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

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

$$\Rightarrow \quad [PRESS_h] \leq 0.95 \times [RESS_{h-1}]$$

Soit :

$$Q^2 = 1 - \frac{PRESS_h}{RESS_{h-1}} \geq 0.05$$

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 $[Q^2_{cum}]_h > 0.5$

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 :

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

- Moyenne des carrés des VIP = 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

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

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
	(Constant)	12070.406	194786.6		.062	.951
	CYLINDRE	-1.936	33.616	-.018	-.058	.955
	PUISSANC	1315.906	613.510	.888	2.145	.047
	VITESSE	-472.507	740.319	-.207	-.638	.532
	POIDS	45.923	100.047	.184	.459	.652
	LONGUEUR	209.653	504.152	.151	.416	.683
	LARGEUR	-505.429	1501.589	-.067	-.337	.741

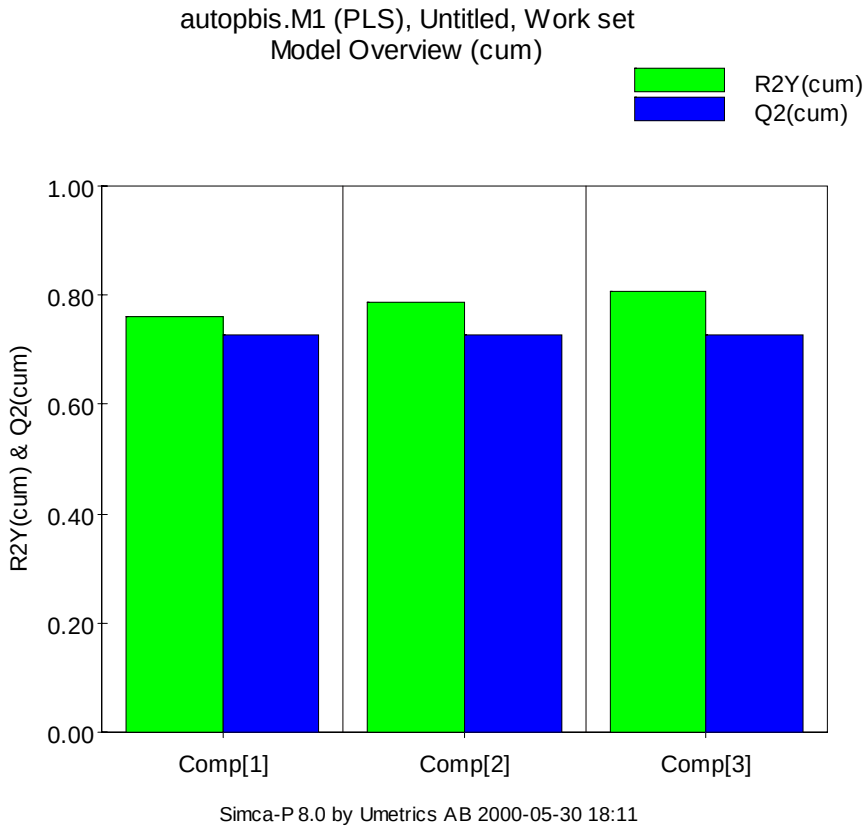
a. Dependent Variable: PRIX

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

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)

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

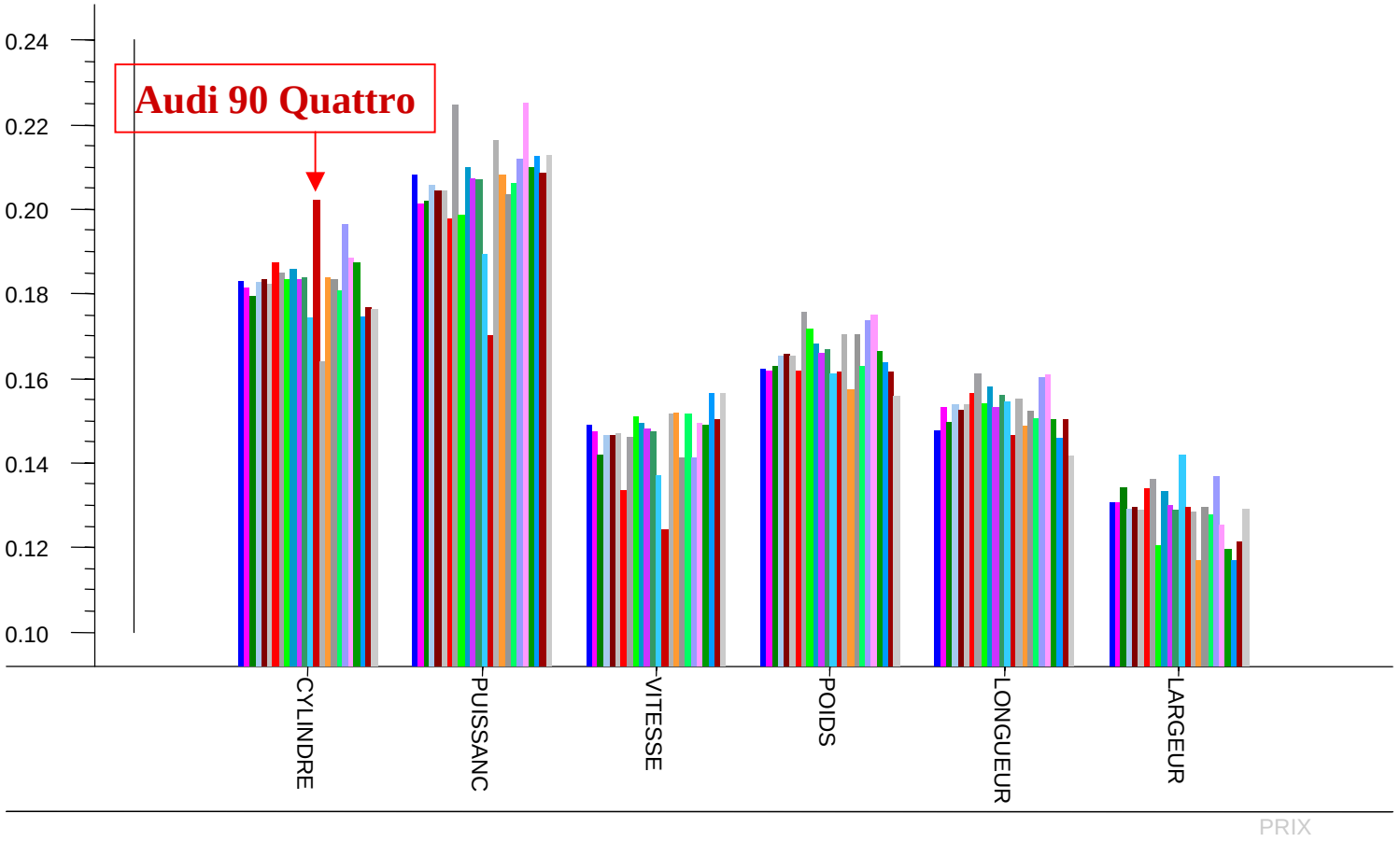
Équation sur les données d'origine (Coeff)

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

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

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

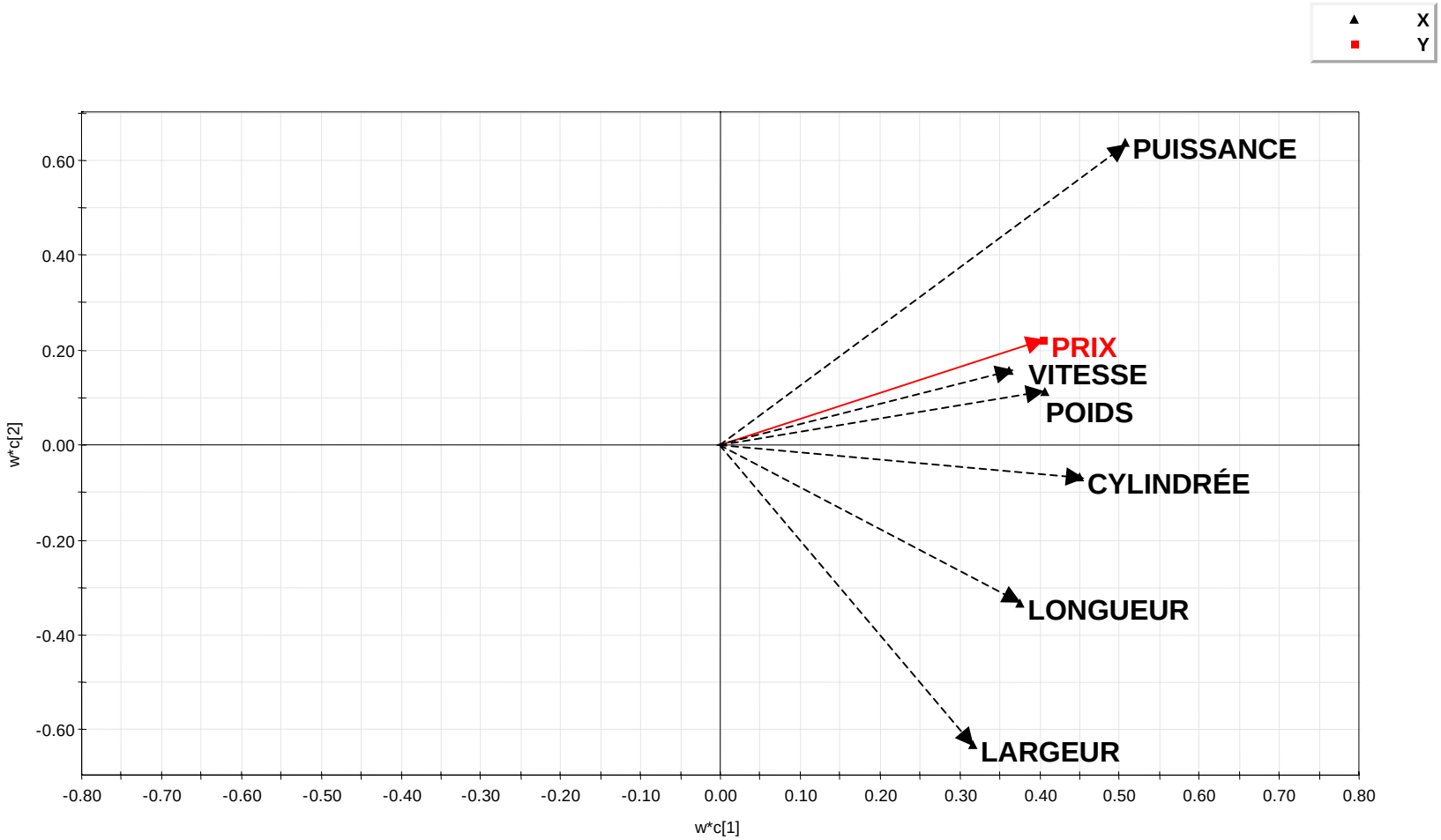
Résultats de la validation croisée sur les coefficients de régression PLS



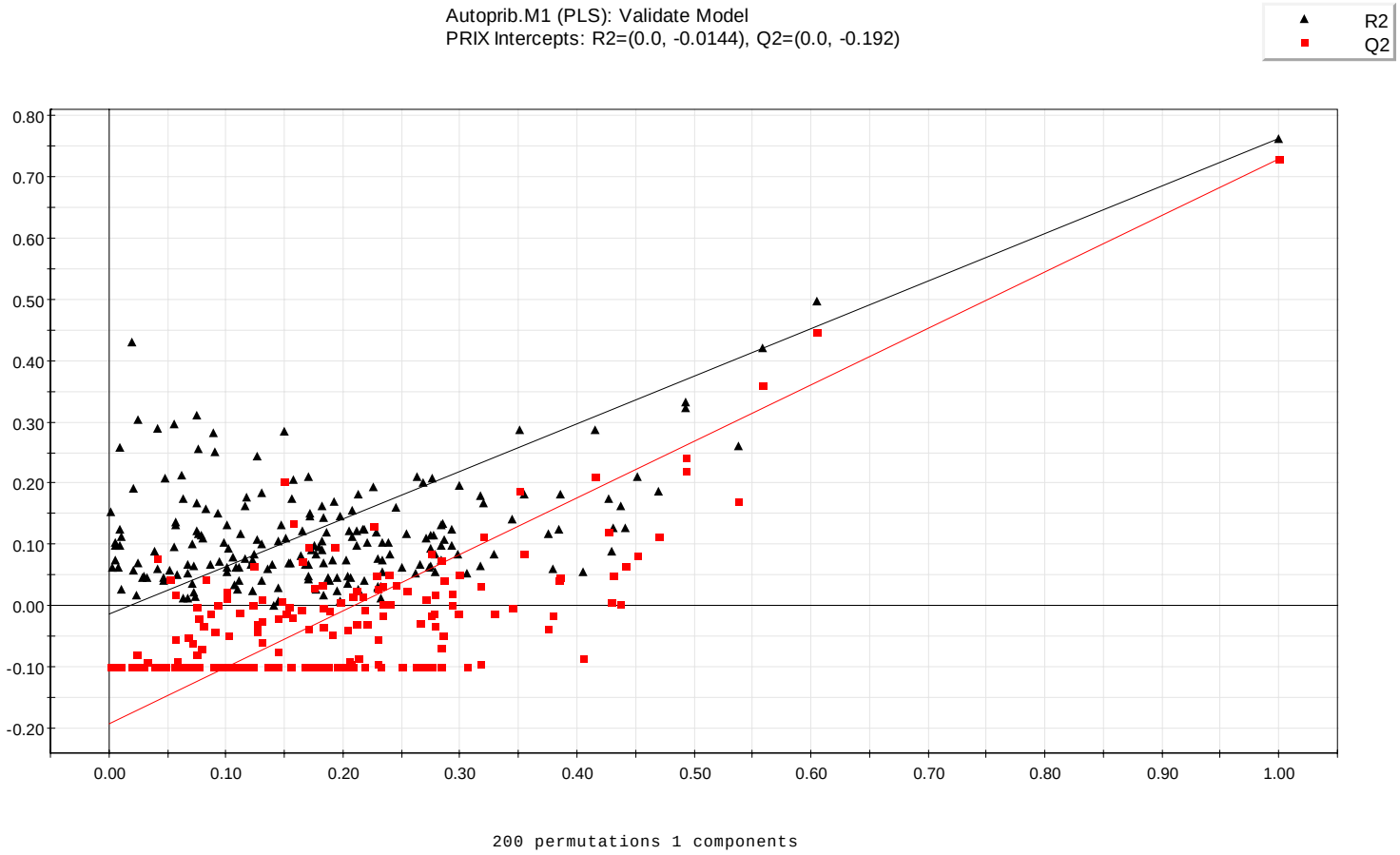
Résultats de la validation croisée sur les coefficients de régression PLS

	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

Carte des variables

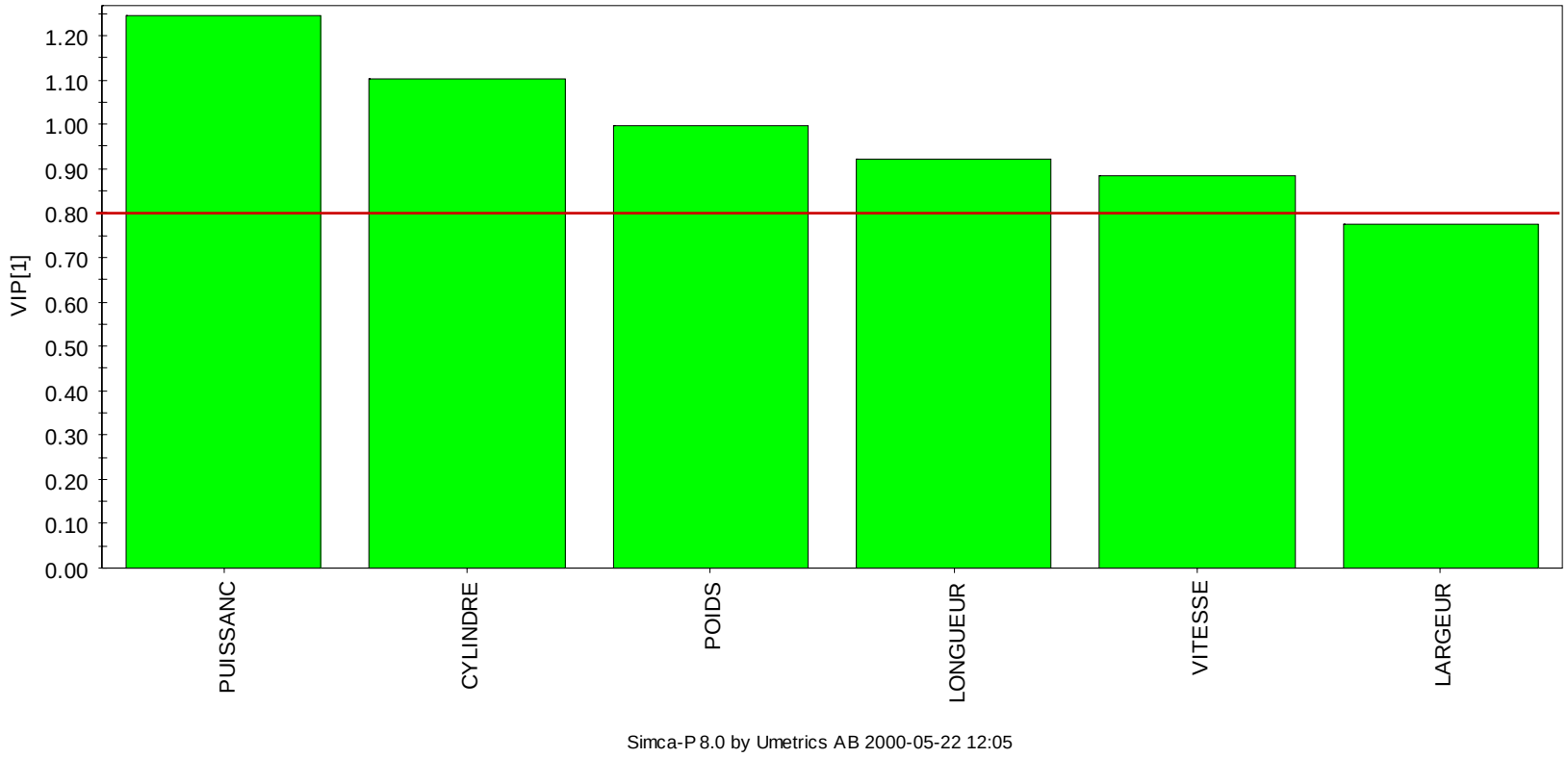


Validation globale

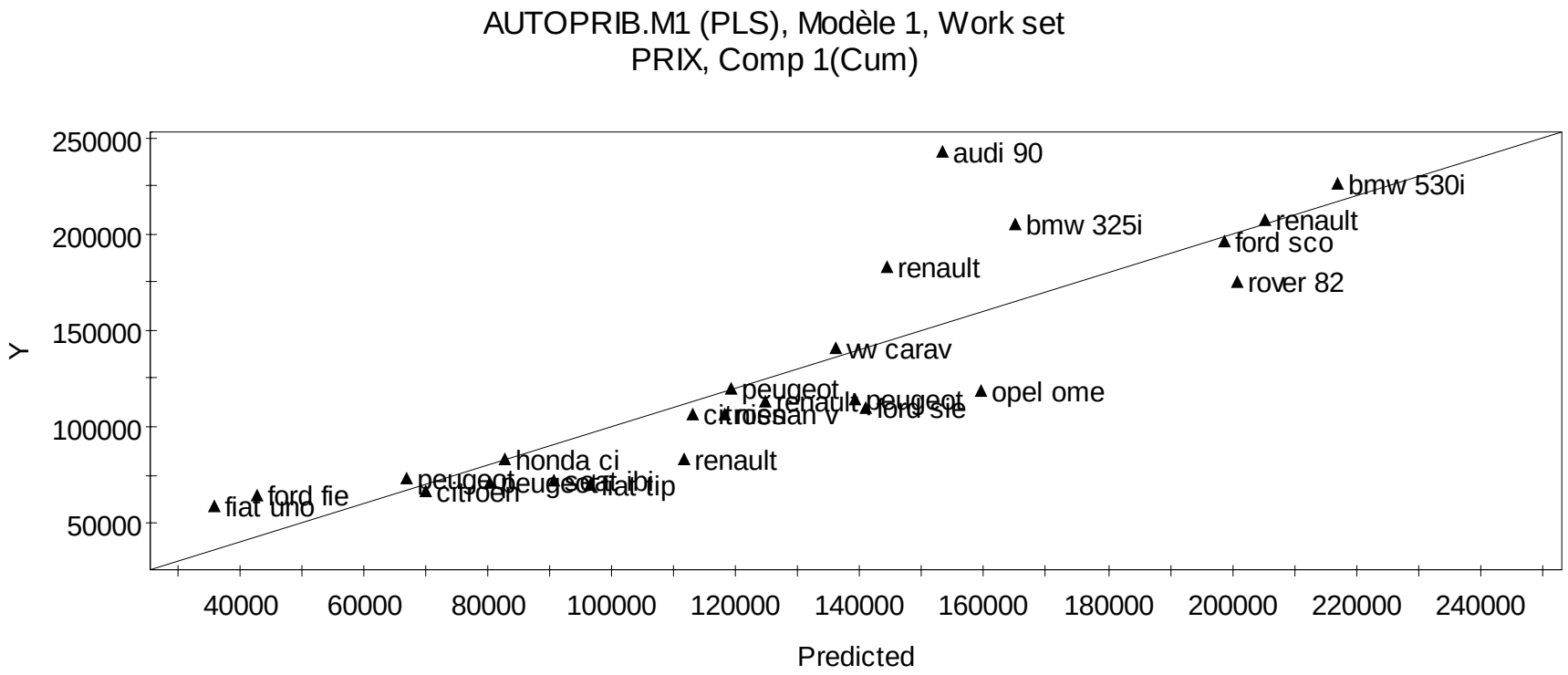


- Abscisse : Corrélation entre Y et Y permuté
- Ordonnée : R2 et Q2 de la régression PLS de Y permuté sur X
- Les droites noire et rouge sont les droites des moindres carrés

Exemple Voitures
Variable Importance in the Projection (1 composante)



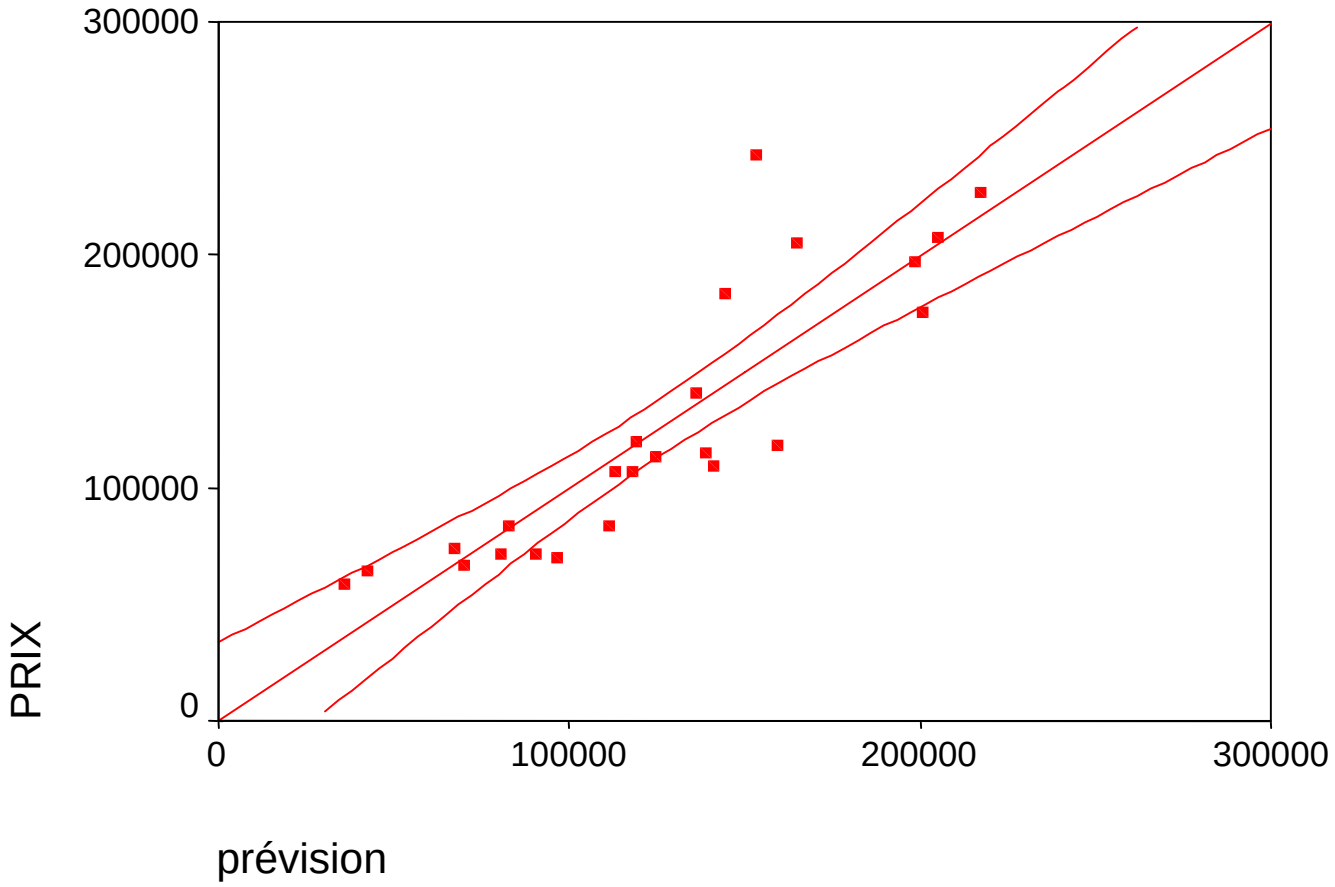
Régression PLS sur les données incomplètes



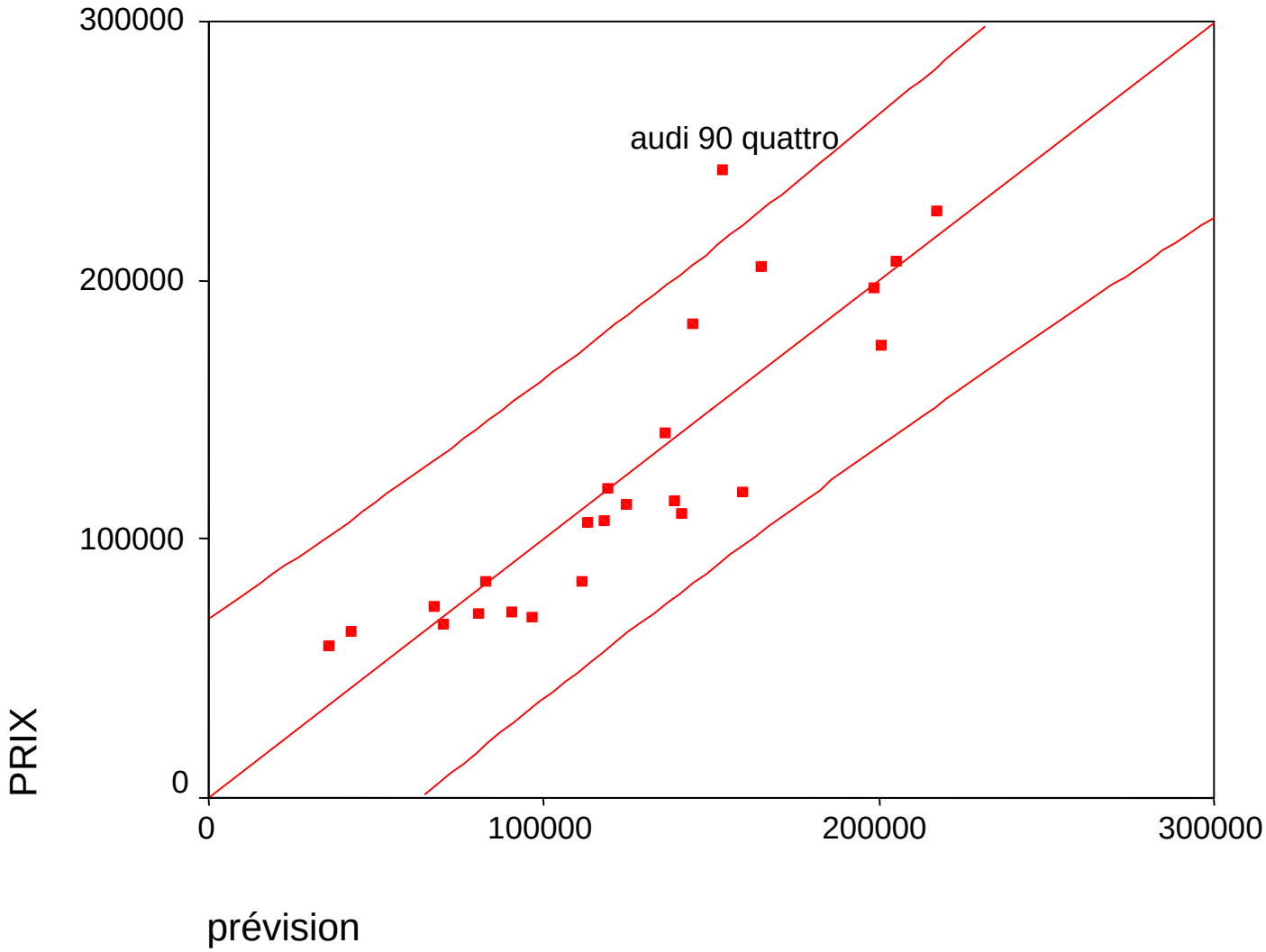
RMSEE=28979

Simca-P 7.01 by Umetri AB 1998-11-23 09:40

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



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



Prédiction du prix de la HONDA CIVIC (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

Régression du Prix sur t_1 :

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

Calcul de tPS_1 pour la HONDA CIVIC :

- Régression : $X_j = p_{1j}t_1 + \text{erreur}, \; j = 1,\ldots, p$

  $\Rightarrow \; p_1 = (p_{11}, \ldots, p_{1p})$

- Régression : $x_i = tPS_{1i}p_1 + \text{erreur}$

  sur les données disponibles; d 'où le calcul de tPS_{1i}

  $\Rightarrow$ tPS_1(Honda Civic) = -1.84262 est l'estimation de t_{1i}

Prédiction du prix de la HONDA CIVIC

- On utilise tPS1 à la place de t_1

  $\Rightarrow$ Prédiction du Prix = 82 644.5 FF

Prédiction du Prix de la HONDA CIVIC : calcul de tPS_1 (Honda Civic)

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

$$x_{Honda} = \begin{pmatrix} ? \\ -.61 \\ -.32 \\ -1.10 \\ -1.23 \\ -.33 \end{pmatrix} \approx tPS_1(Honda) \times \begin{bmatrix} 0.48 \\ 0.45 \\ 0.37 \\ 0.39 \\ 0.39 \\ 0.36 \end{bmatrix} \Rightarrow$$

tPS_1(Honda) = -1.84262

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

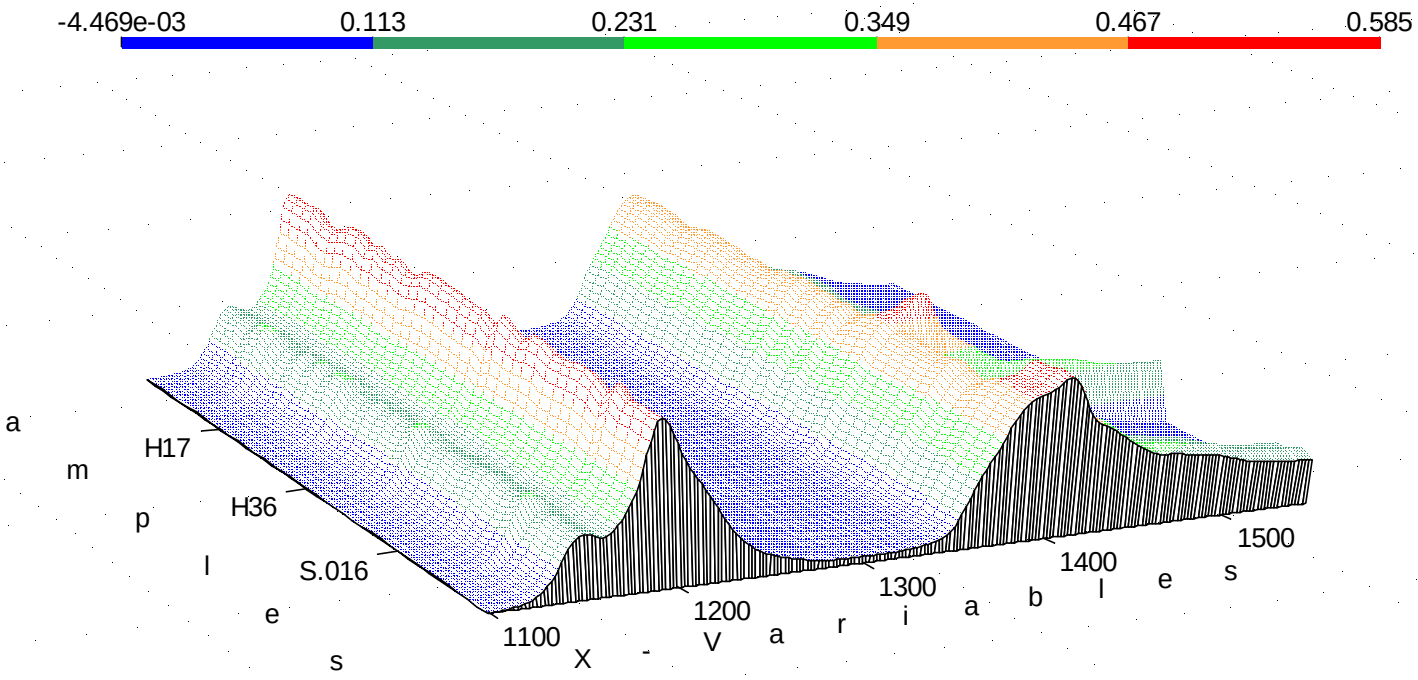
Les données :

- 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)

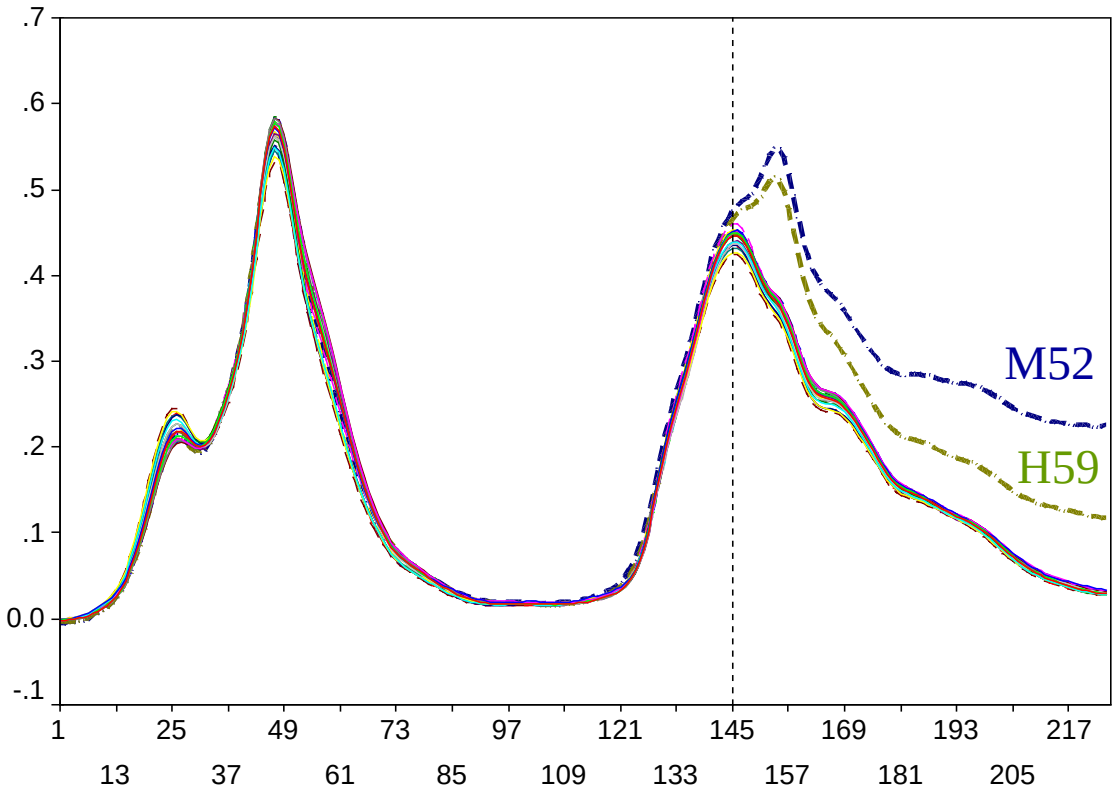
Cas UOP Guided Wave
Visualisation des X



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

Cas UOP Guided Wave

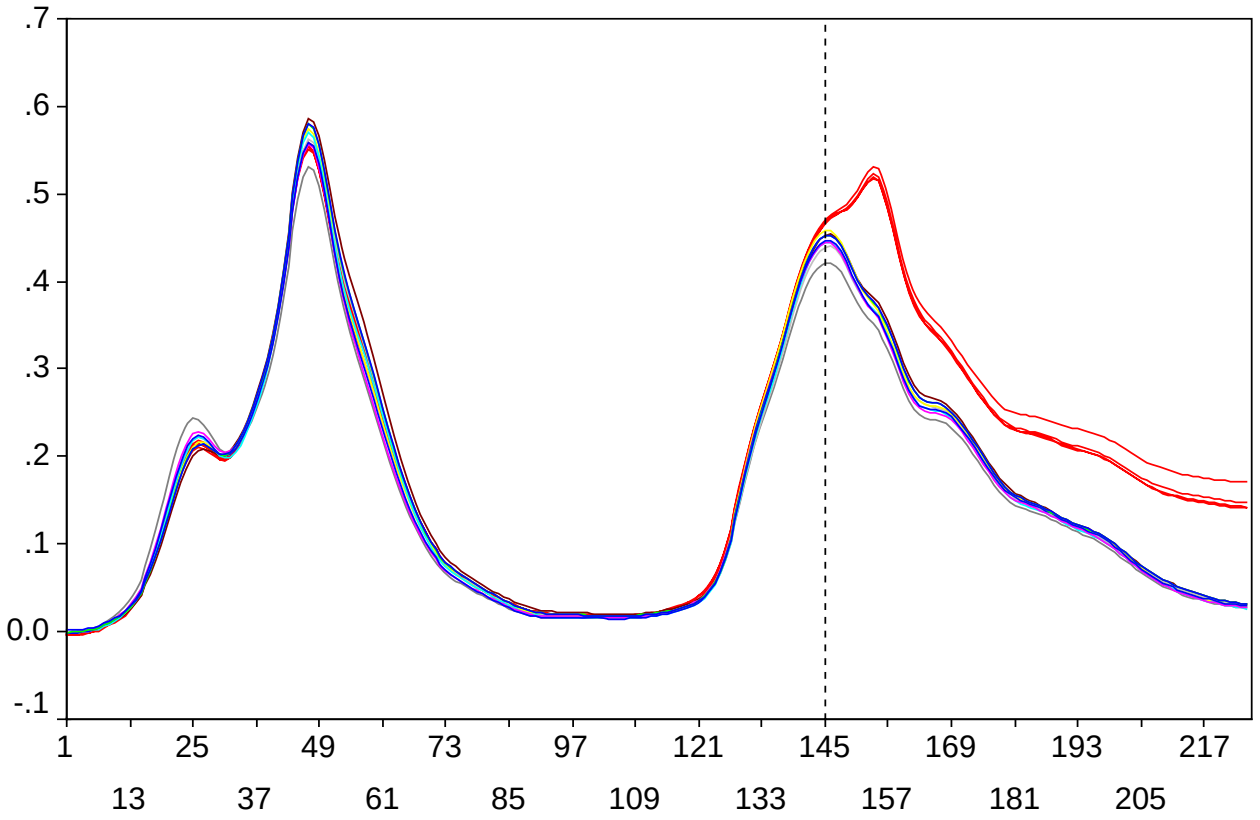
Visualisation des X : Données de calibration



Les échantillons M52 et H59 contiennent de l ’alcool

Cas UOP Guided Wave

Visualisation des X : Données de validation



Numéro de la longueur d'onde

Les échantillons avec alcool sont en rouge

Régression PLS1 : les résultats

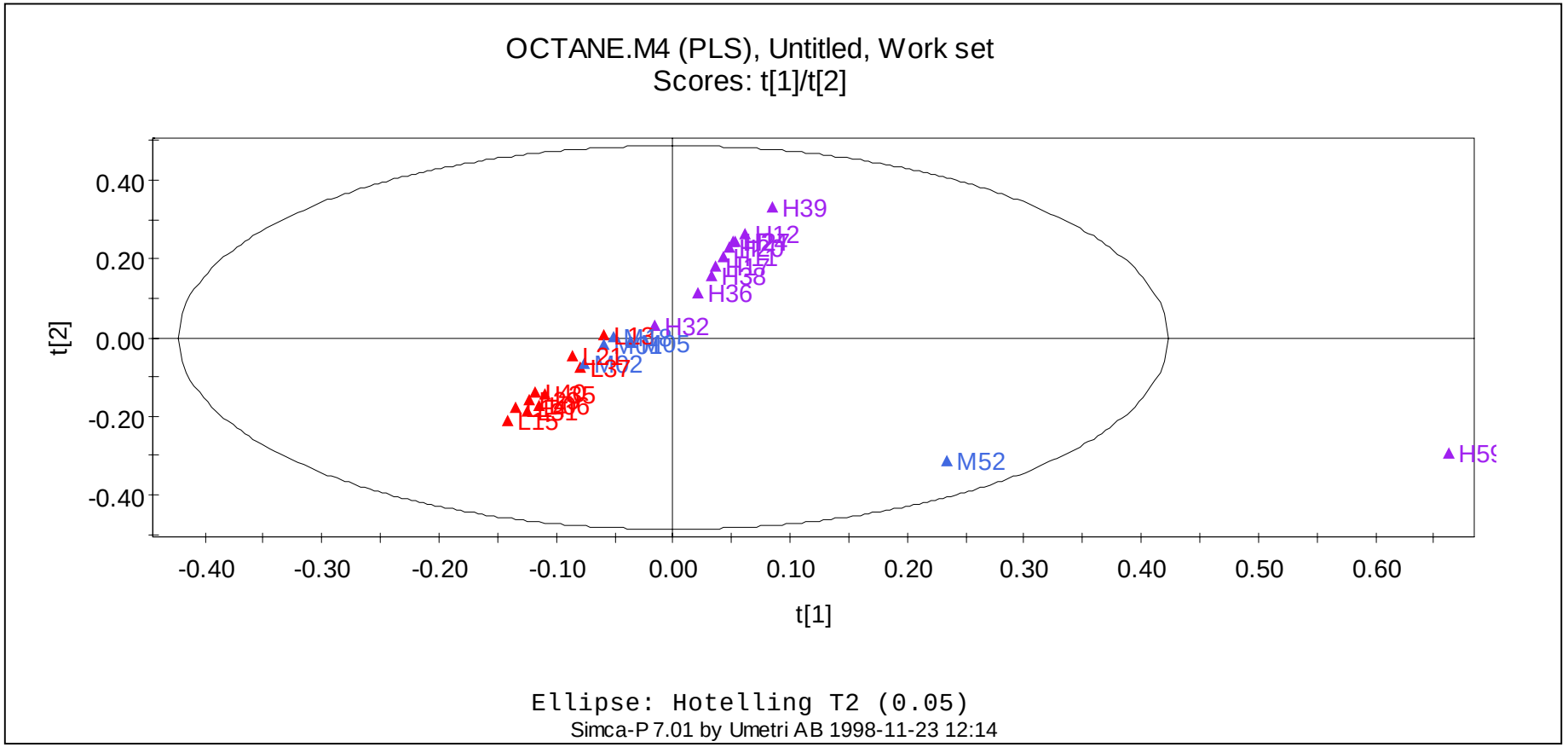
- Données de spectroscopie

  Les données sont centrées, mais non réduites

- Validation croisée :

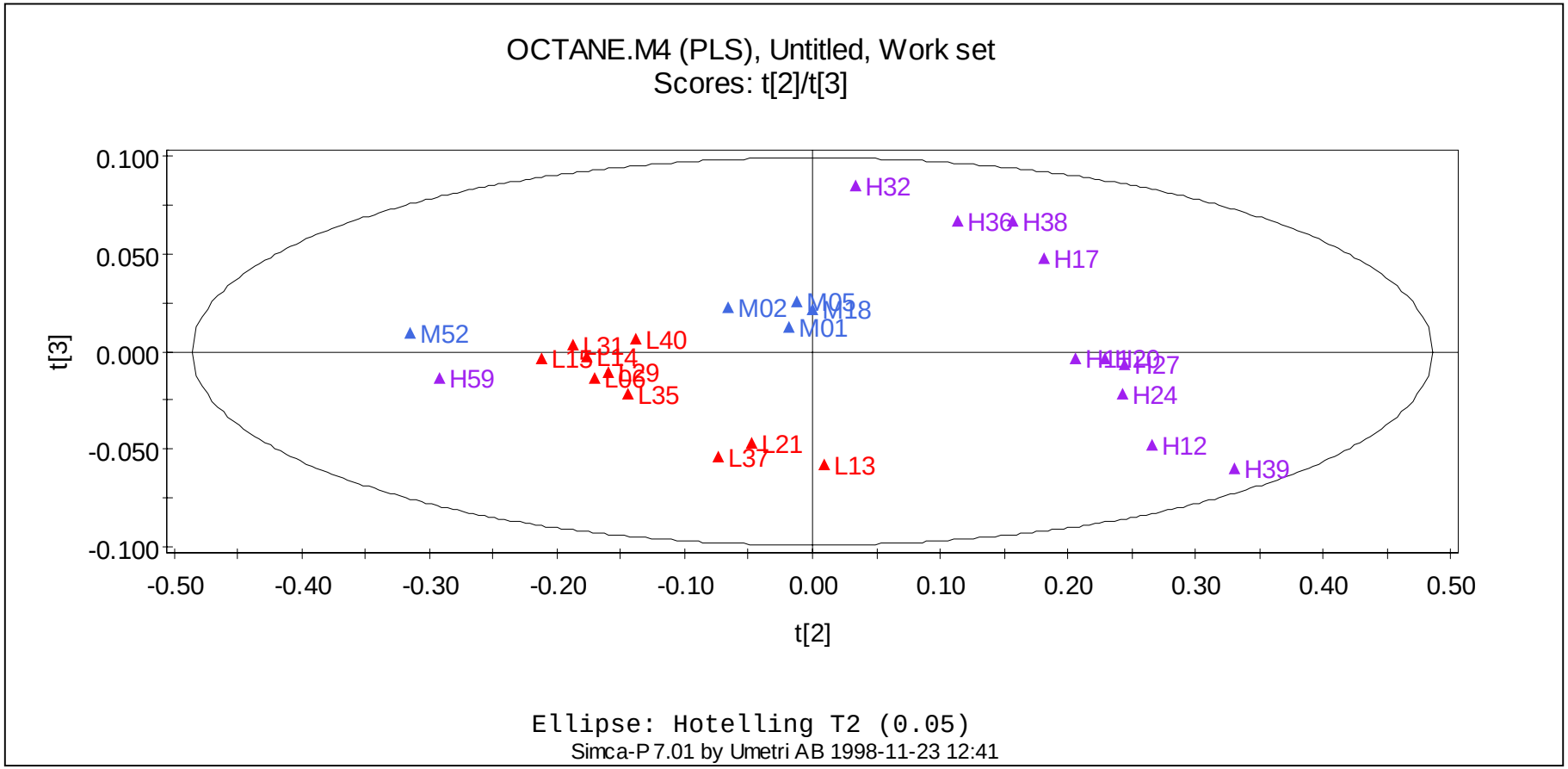
  3 composantes PLS

UOP Guided Wave : Les composantes PLS



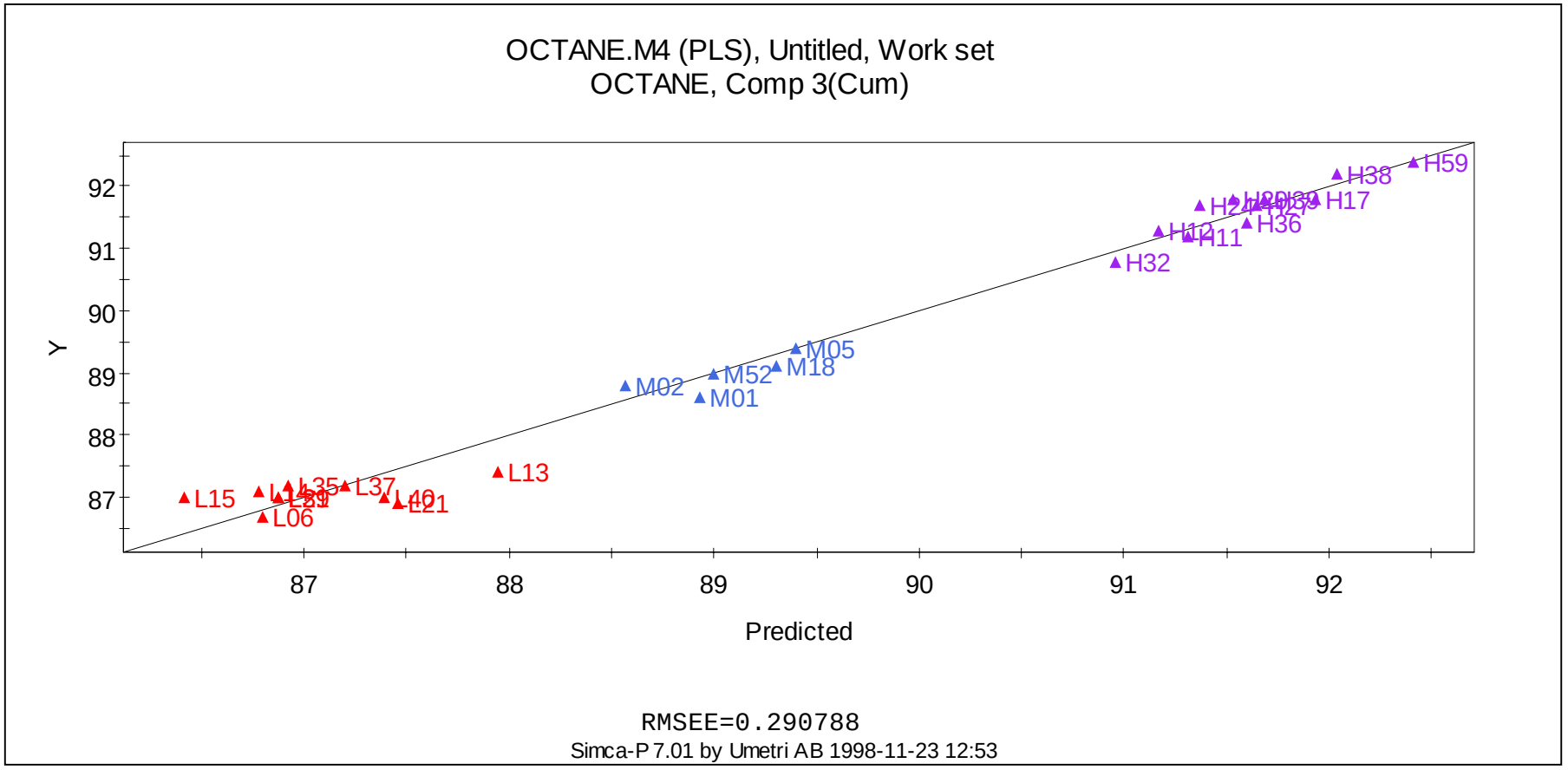
- Indice d ’octane : L = Low, M = Medium, H = High
- Les échantillons M52 et H59 contiennent de l’alcool

UOP Guided Wave : les composantes PLS

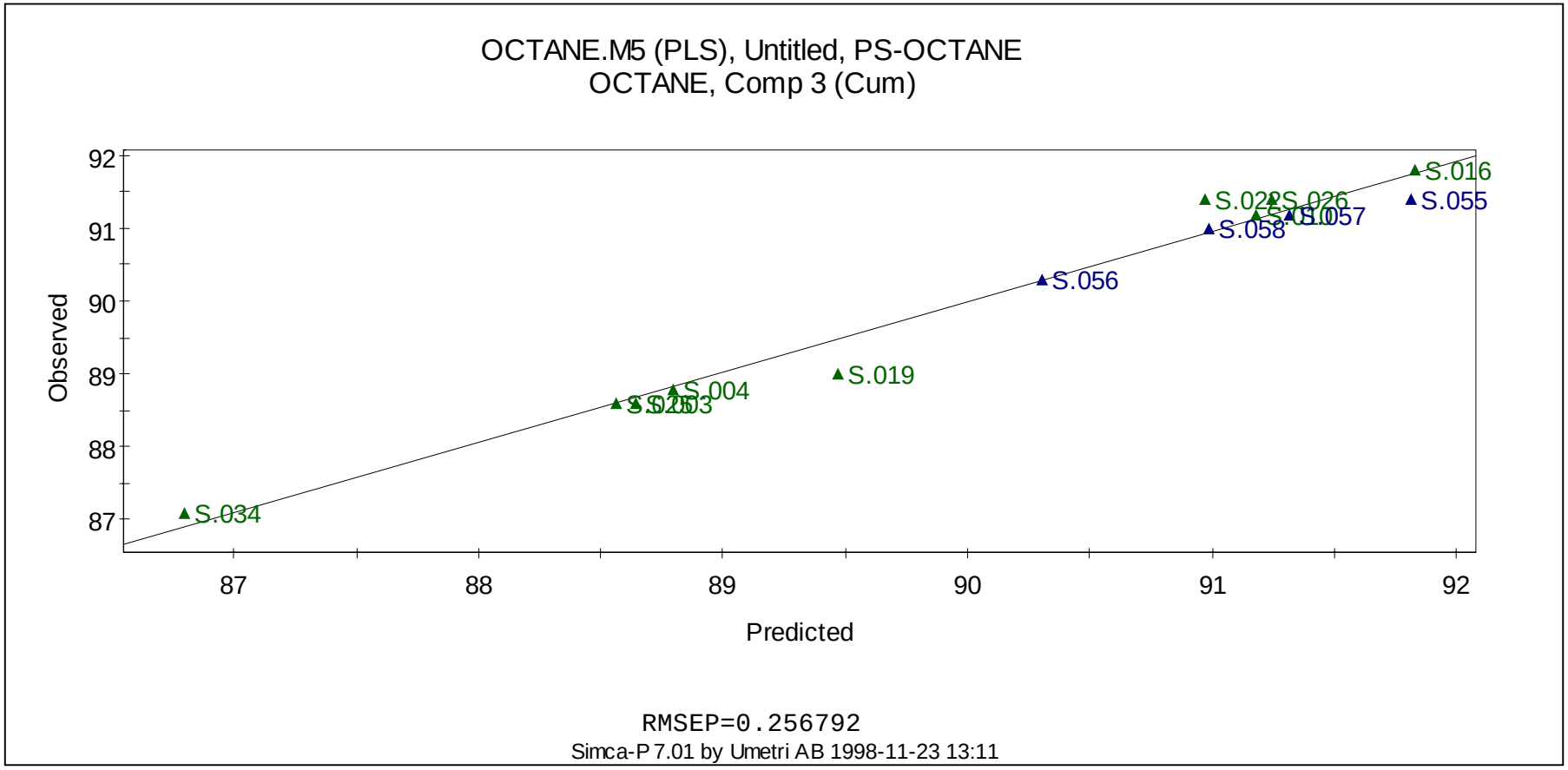


Indice d ’octane : L = Low, M = Medium, H = High

Cas UOP Guided Wave : Prévision
Données de calibration



Cas UOP Guided Wave : Prévision
Données de validation



Présence d ’alcool : OUI / NON

II.2 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.**

La régression PLS2 : une idée de l'algorithme

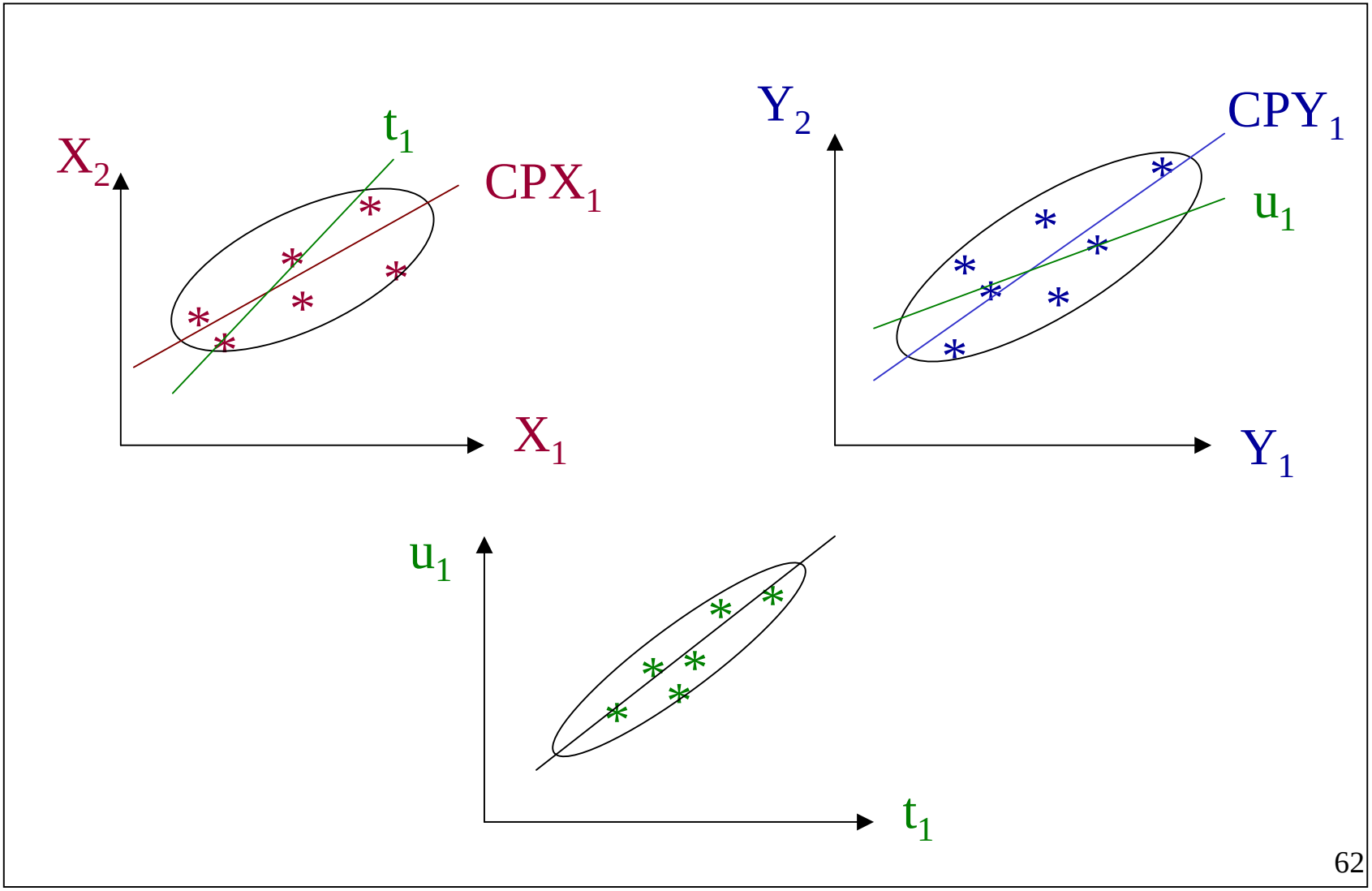
Etape 1 : Recherche de m composantes orthogonales $t_h = Xa_h$ et m composantes $u_h = Yb_h$ bien corrélées entre elles et explicatives de leur propre groupe.

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.

Objectif de l'étape 1 de la régression PLS2



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

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

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

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.

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 :

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

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

Régression PLS2

Exemple 1: 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

Cas Dégustation de thé

- **Bloc X**

  Variables indicatrices des modalités

  de Température, Sucré, Force et Citron

- **Bloc Y**

  Les classements des sujets

Cas Dégustation de thé
Résultats de la régression PLS

- **Validation croisée** :

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

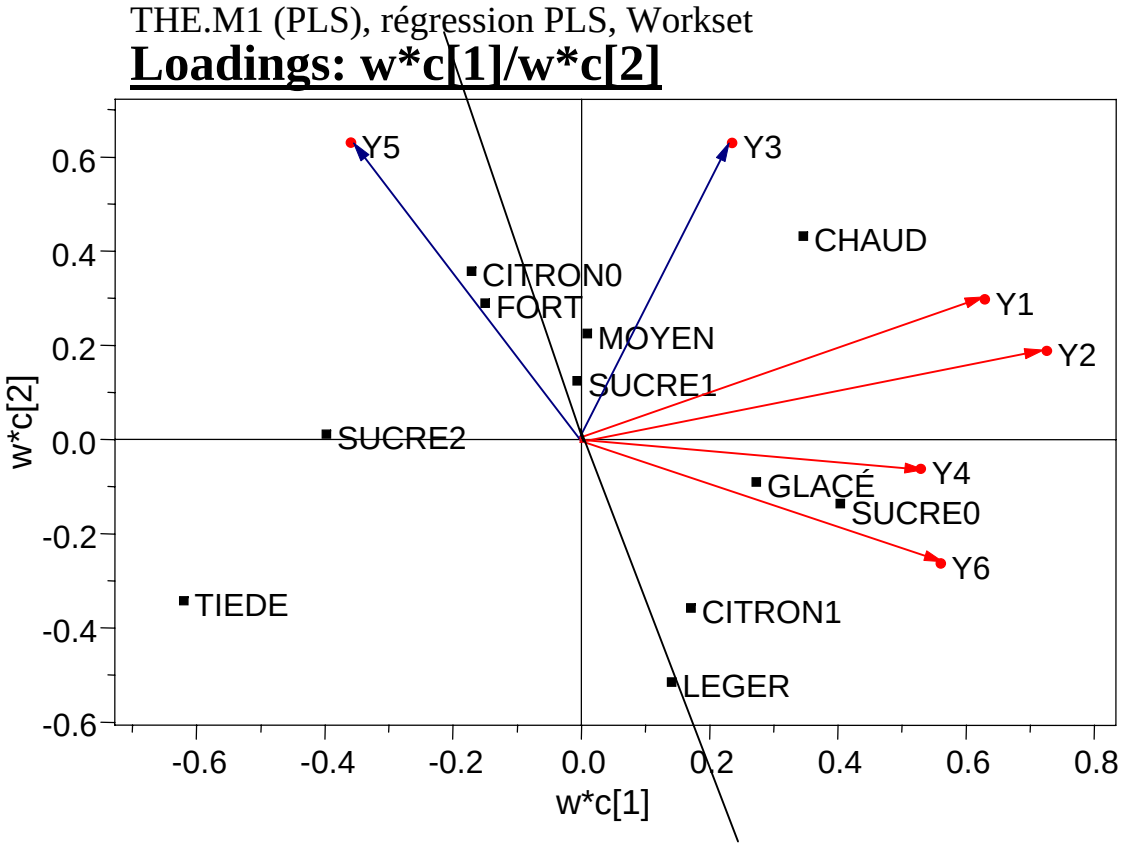
- **Équation de régression de Y_k sur $t_1, \ldots, t_h$** :

  $Y_k = c_{1k}t_1 + c_{2k}t_2 + c_{3k}t_3 + c_{4k}t_4 + \text{résidu}$

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

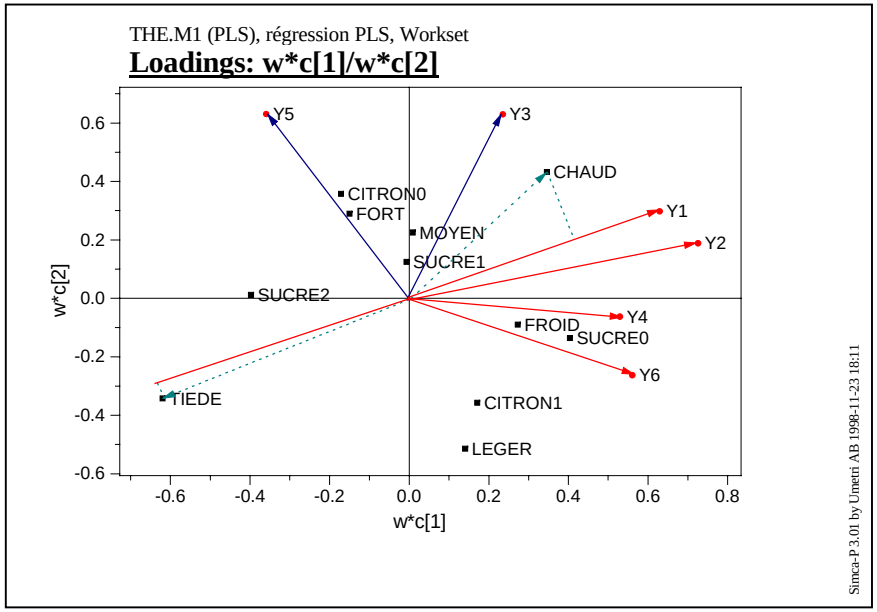
Cas Dégustation de thé

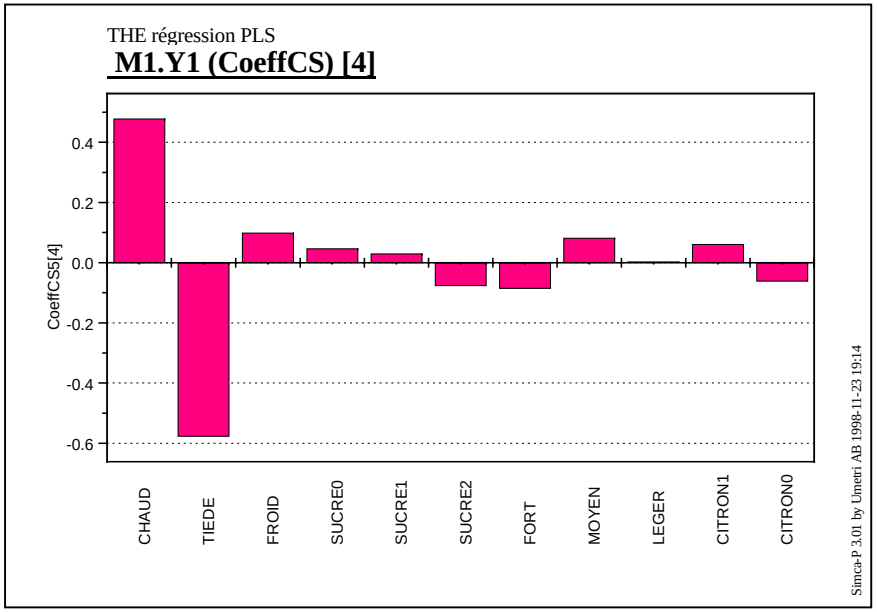
Carte des variables



Cas dégustation de thé

Visualisation de la régression PLS de Y_1 sur X



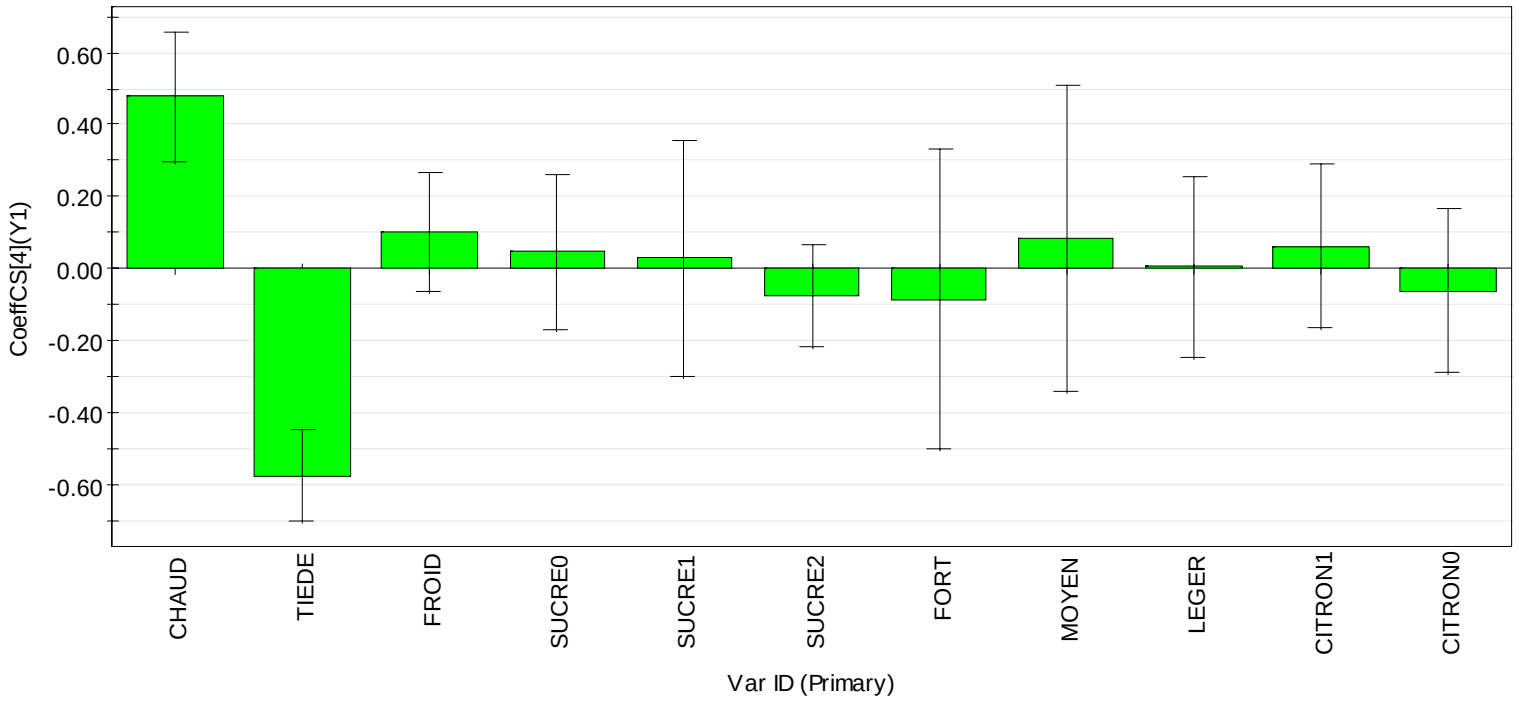


Règle d'interprétation:

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.

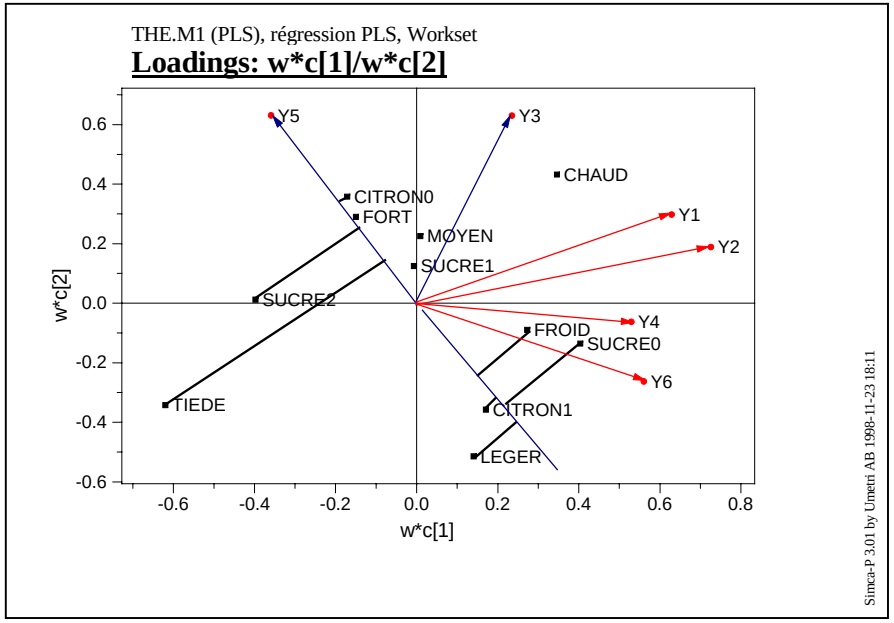
Le juge 1 aime son thé chaud et rejette le thé tiède

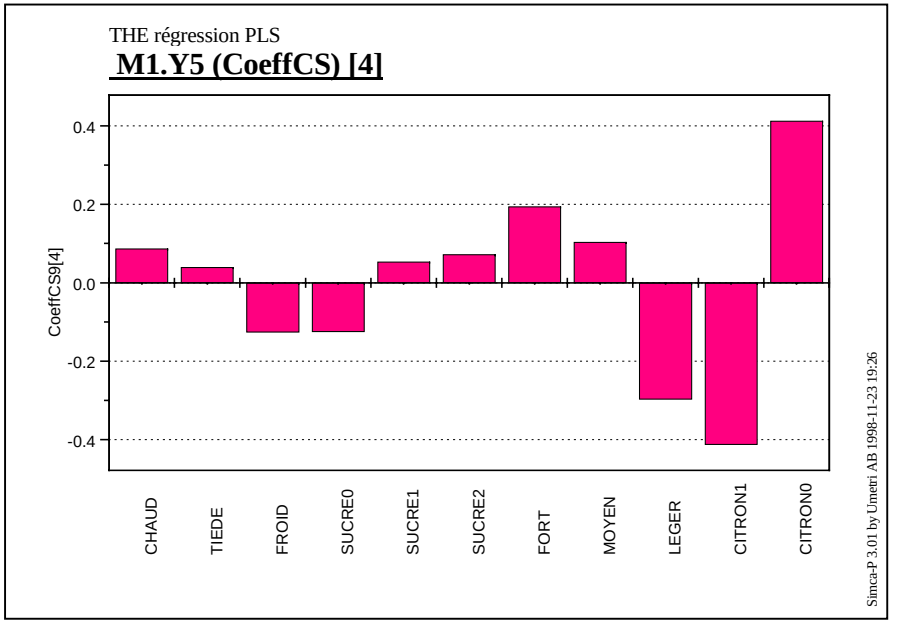
Validation du modèle pour le juge 1



Cas dégustation de thé

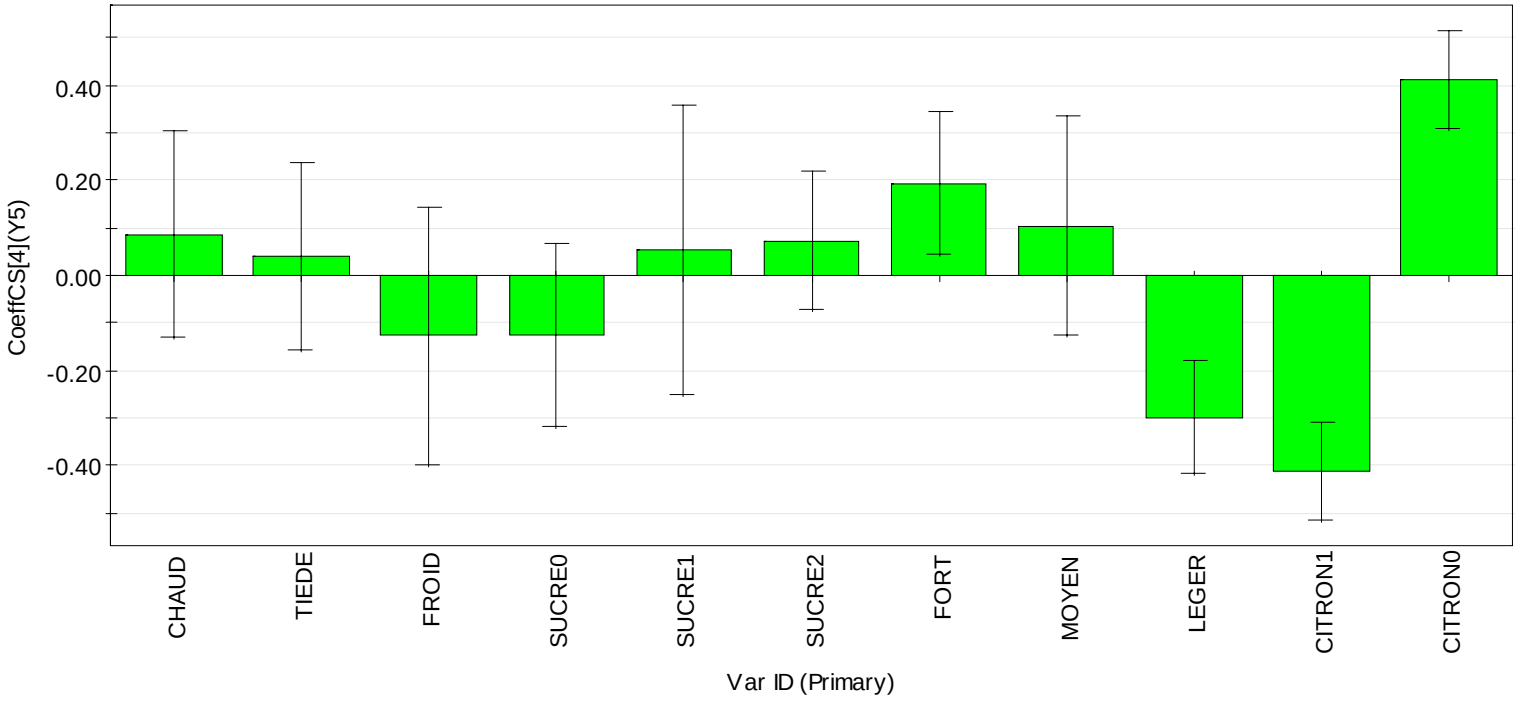
Visualisation de la régression PLS de Y_5 sur X



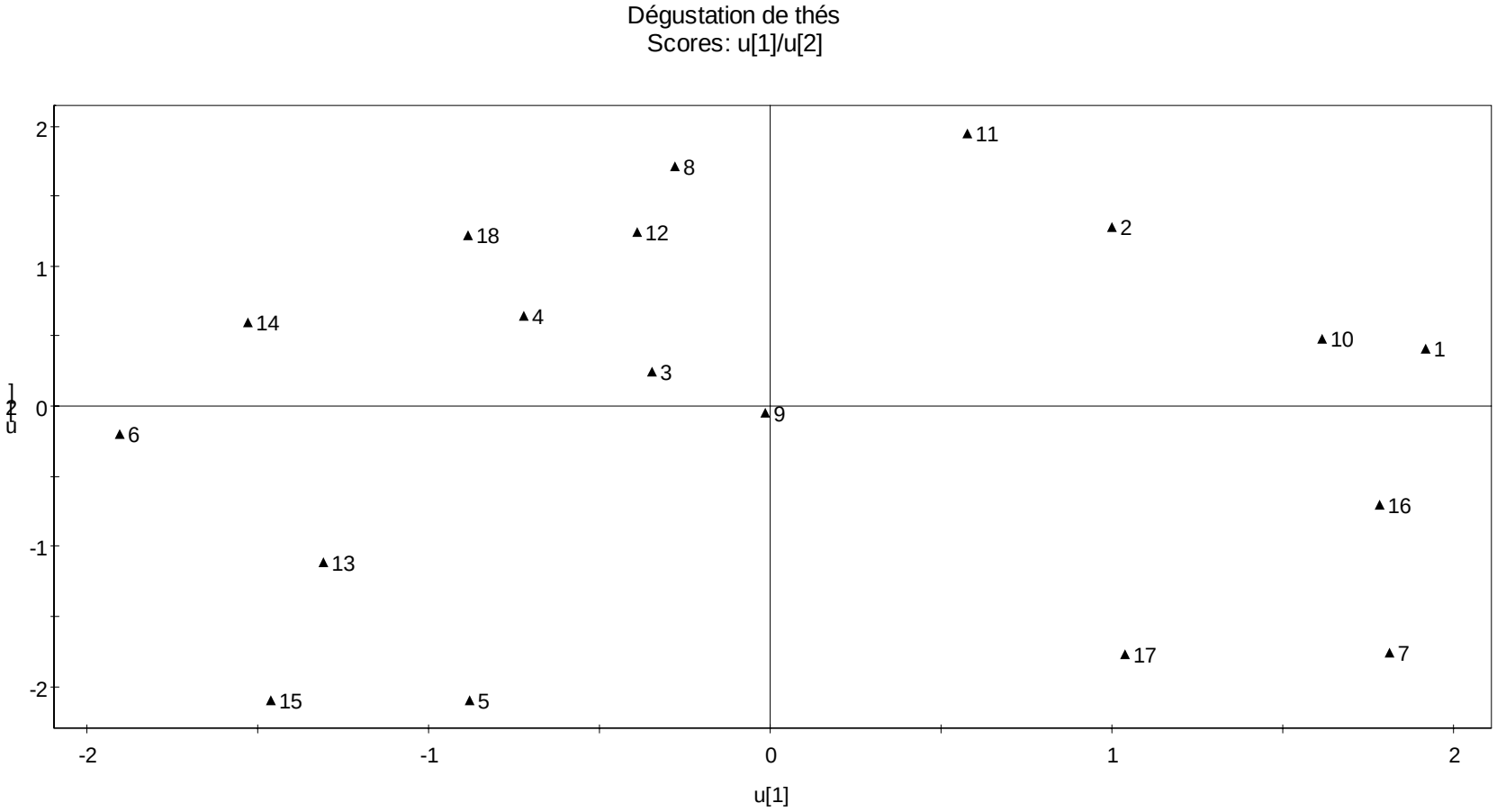


Le juge 5 préfère son thé sans citron, fort;
il est indifférent au thé tiède; il rejette le thé léger, avec du citron.

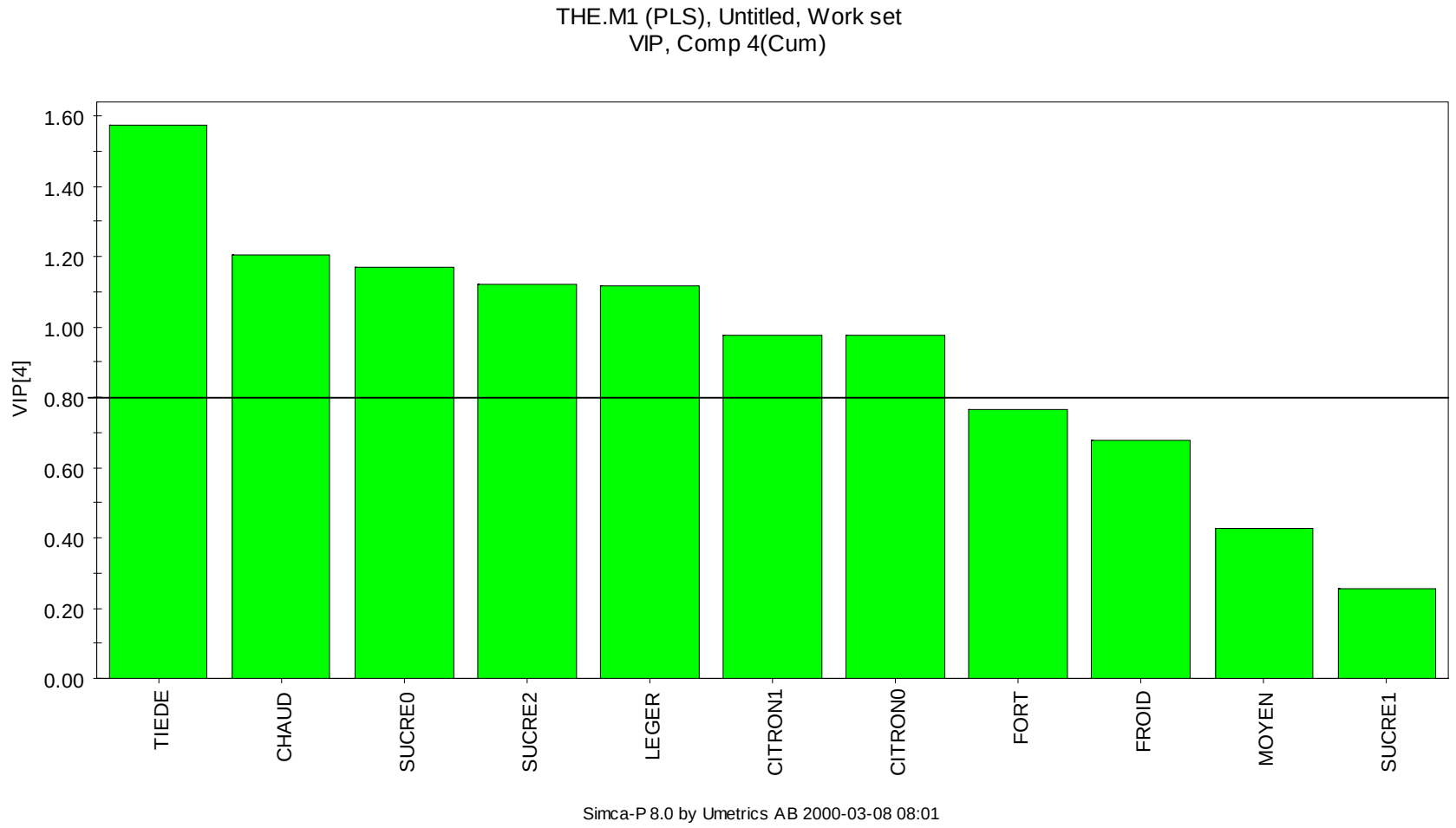
Validation du modèle pour le juge 5



Carte des produits dans l'espace des juges



Variable Importance in the Projection (VIP)



III. 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 PLS de Y sur X**

Analyse discriminante PLS : exemple

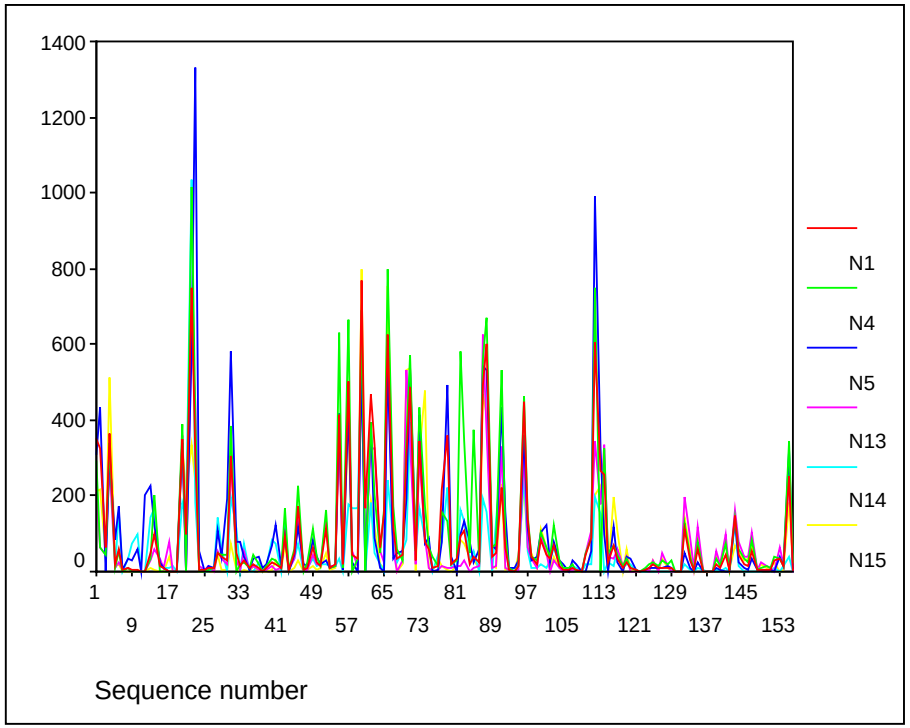
Les données

- 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

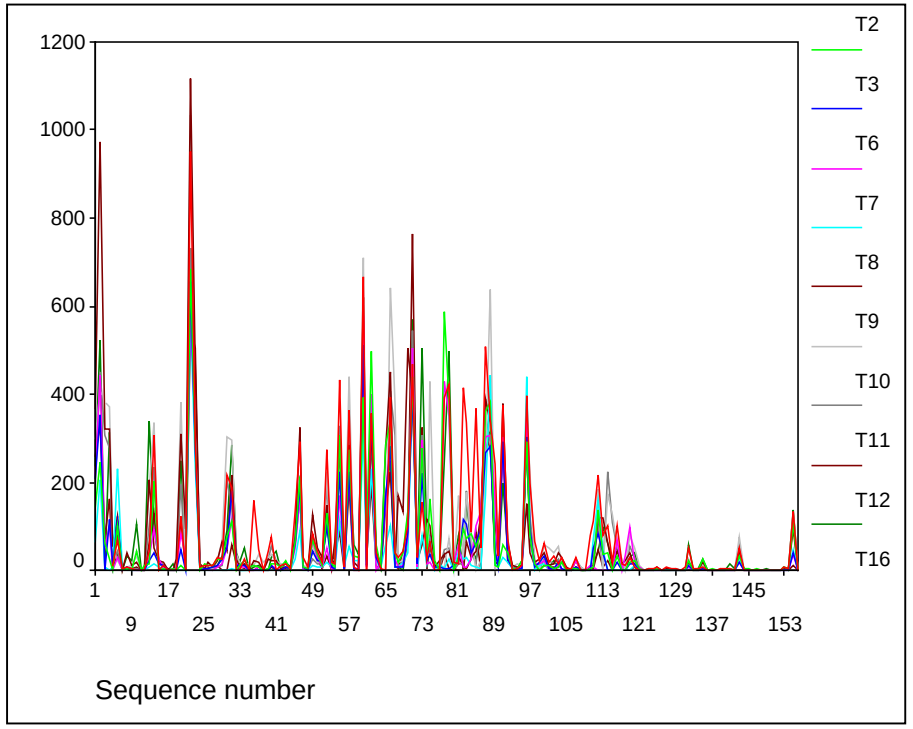
Article:

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
Profils métaboliques des biopsies

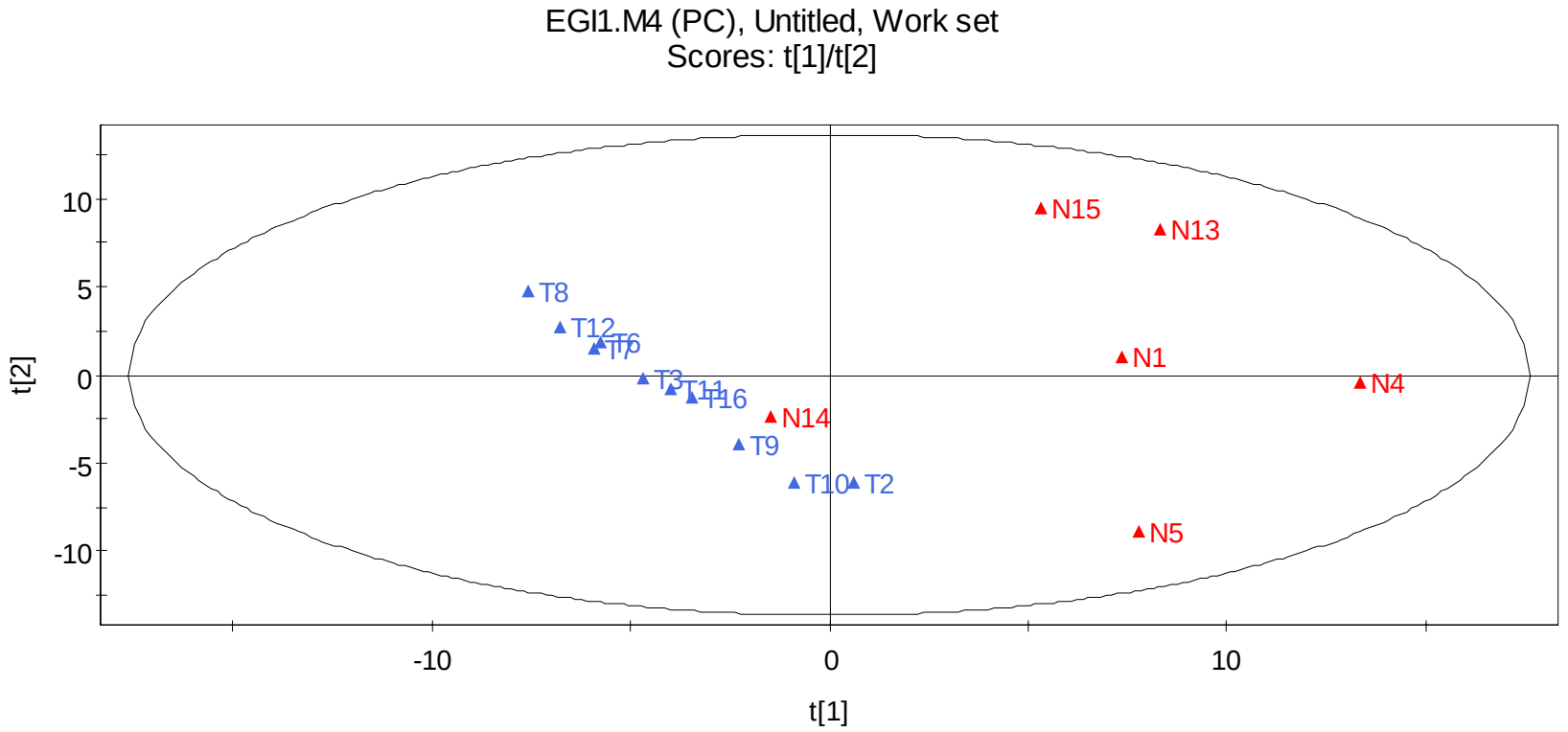


Tumeurs bénignes

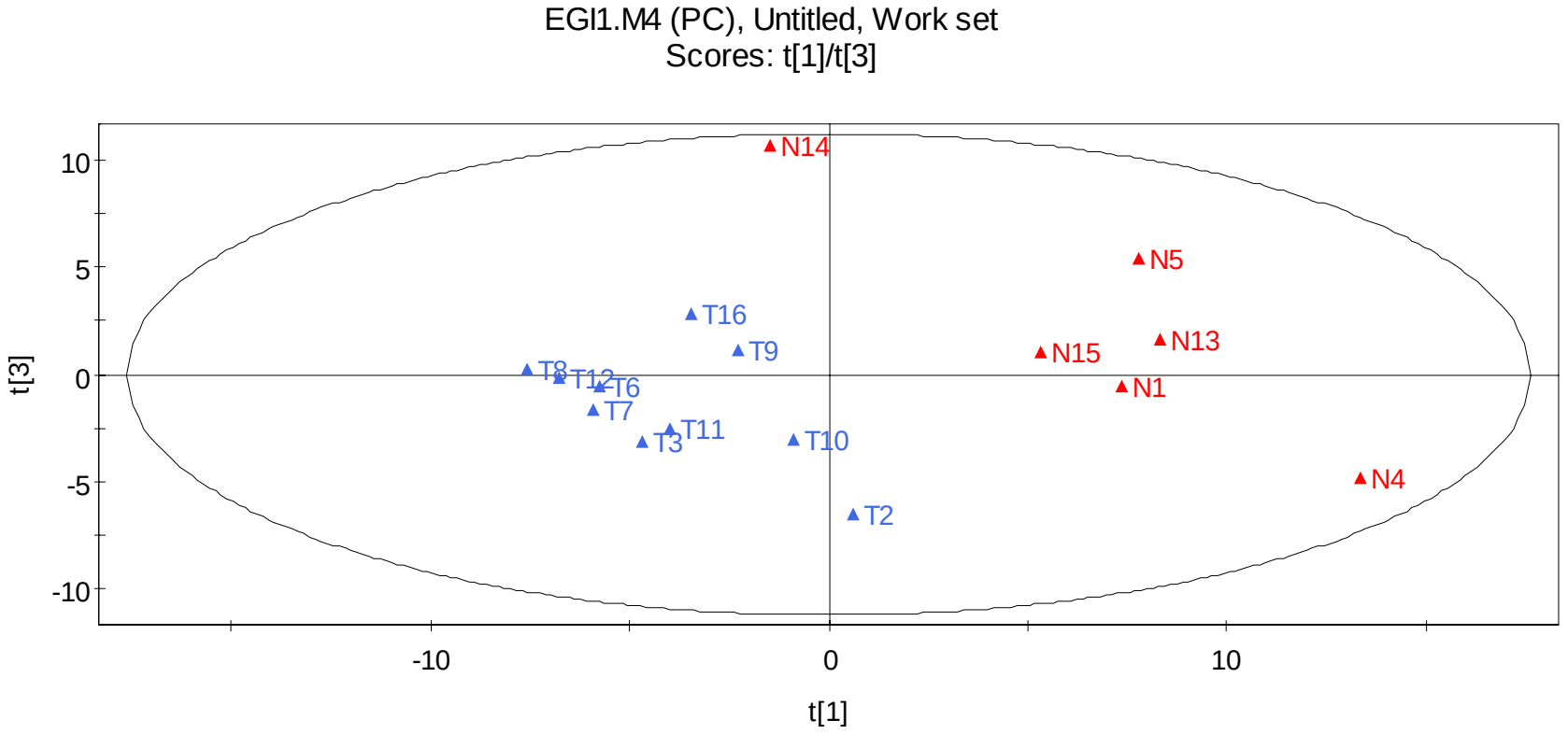


Tumeurs malignes

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



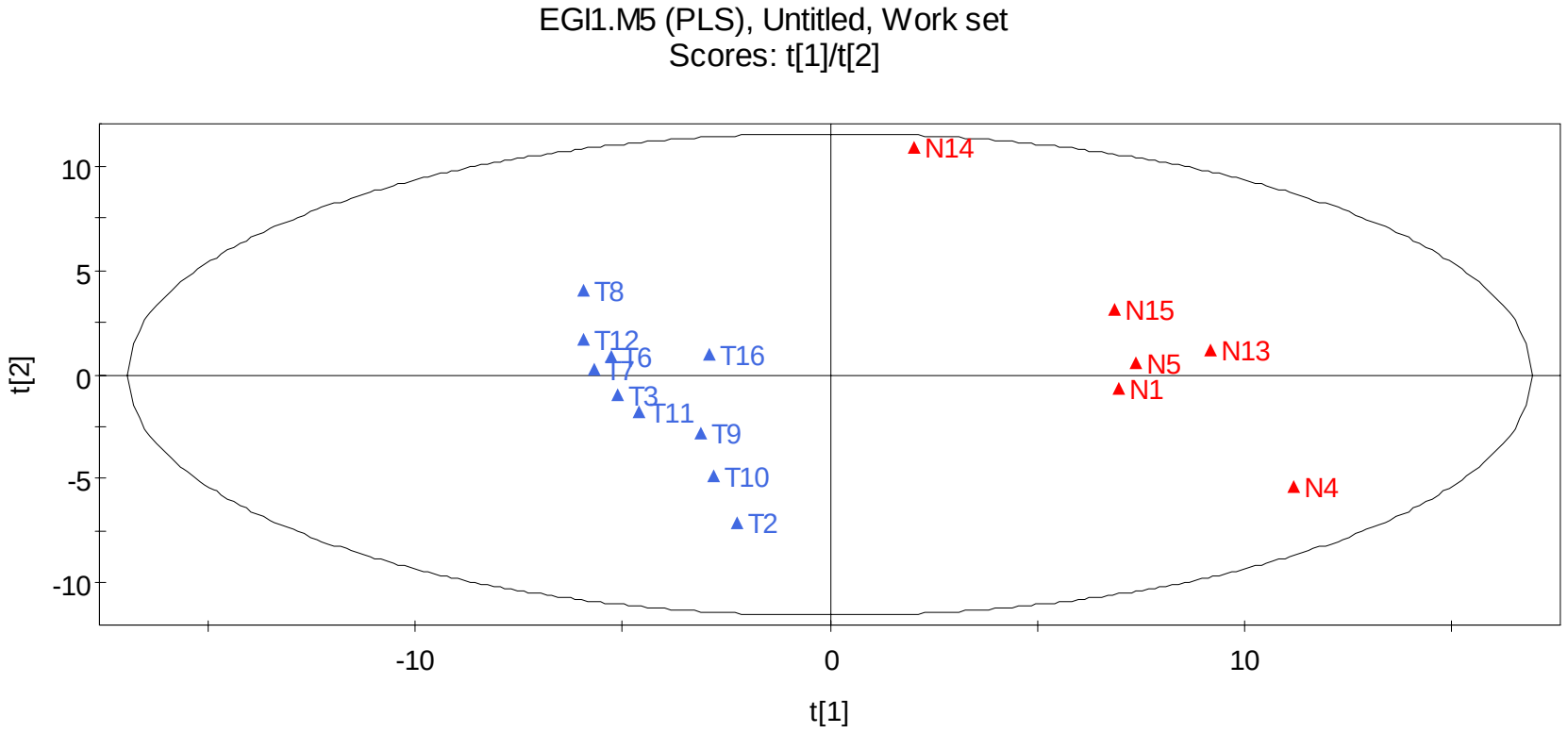
Analyse en composantes principales des 16 biopsies
Composantes principales 1 et 3



Ellipse: Hotelling T2 (0.05)

Simca-P 7.01 by Umetri AB 1998-11-24 15:19

Analyse discriminante PLS
Composantes PLS 1 et 2



Ellipse: Hotelling T2 (0.05)

Simca-P 7.01 by Umetri AB 1998-11-24 15:22

IV. 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 trois algorithmes

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

Régression logistique ordinale

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

PROB(Y ≤ i) =

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

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

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

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**.

Algorithme 1 : La régression logistique PLS

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

Le nombre m est obtenu par validation croisée.

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

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

Régression logistique PLS
Étape 1

1. Régression logistique de y sur chaque x_j :

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

2. Normalisation du vecteur $a_1 = (a_{11},\ldots,a_{1k})$

3. Régression logistique de y sur $T_1 = Xa_1$
   exprimée en fonction des X

4. Calcul du résidu X_1 de la régression de X sur T_1

Régression logistique PLS
Étape 2

1. Régression logistique de y sur T_1 et chaque résidu x_{1j} :

   $\Rightarrow$ les coefficients de régression b_{2j}
2. Normalisation du vecteur $b_2 = (b_{21},\ldots,b_{2k})$
3. Calcul de a_2 tel que : $T_2 = X_1 b_2 = X a_2$
4. Régression logistique de y sur $T_1 = X a_1$ et $T_2 = X a_2$ exprimée en fonction des X
5. Calcul du résidu X_2 de la régression de X sur T_1 , T_2

Régression logistique PLS
Choix du nombre de composantes

- On procède de la même manière pour les autres étapes.
- On choisit le nombre de composantes par validation croisée : la composante h est retenue si

$$[\chi^2_{Pearson}(\text{validation croisée, étape h})]^{1/2} \leq 0.95 \times [\chi^2_{Pearson}(\text{substitution, étape h-1})]^{1/2}$$

Soit :

$$Q^2 = 1 - \frac{\chi^2_{\text{validation croisée, étape h}}}{\chi^2_{\text{substitution, étape h-1}}} \geq 0.0975$$

Régression logistique PLS
Résultats de l'algorithme

- La température de 1924 est supposée inconnue.
- La régression logistique PLS de Y sur X a conduit à deux composantes PLS T_1 et T_2 :

$$T_1 = 0.57\times Température + 0.63\times Soleil + 0.41\times Chaleur - 0.34\times Pluie$$

$$T_2 = -0.14\times Température + 0.45\times Soleil - 0.69\times Chaleur - 0.52\times Pluie$$

Régression logistique ordinale sur T_1, T_2
Résultats SAS

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCP1	1	-2.5490	0.8768	8.4507	0.0036
INTERCP2	1	2.1349	0.8955	5.6837	0.0171
T1	1	3.0797	0.8350	13.6032	0.0002
T2	1	1.4148	0.8849	2.5563	0.1099

TABLEAU CROISANT QUALITÉ OBSERVÉE ET PRÉDITE

QUALITÉ PRÉDICTION

Effectif	1	2	3	Total
1	9	2	0	11
2	1	9	1	11
3	0	1	11	12
Total	10	12	12	34

Résultat :

5 années mal classées

Régression logistique PLS
Le modèle

Prob (Y≤ i)

$$= \frac{e^{-2.55\times Bon+2.14\times Moyen+3.08\times T_1+1.42\times T_2}}{1+e^{-2.55\times Bon+2.14\times Moyen+3.08\times T_1+1.42\times T_2}}$$

$$= \frac{e^{-2.55\times Bon+2.14\times Moyen+1.57\times Temp.+2.73\times Soleil+0.26 Chaleur-1.77\times Pluie}}{1+e^{-2.55\times Bon+2.14\times Moyen+1.57\times Temp.+2.73\times Soleil+0.26\times Chaleur-1.77\times Pluie}}$$

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.

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).
- $t_1 = 0.55 \times \textit{Température} + 0.55 \times \textit{Soleil} + 0.48 \times \textit{Chaleur} - 0.40 \times \textit{Pluie}$
- Pour l'année 1924 :

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

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 (Test for larger residuals than minimum)
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 PREV

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

Régression logistique sur composantes PLS
Le modèle

Prob (Y≤ i)

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

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

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

Conclusion 2:
Le modèle linéaire généralisé PLS

- Le modèle linéaire généralisé 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.

Algorithme 3 (données groupées)
Régression PLS du logit de la variable de réponse sur les prédicteurs

Exemple : Job satisfaction (Zelterman, 1999)

- 9949 employees in the ‘ craft ’ job within a company
- Response : Satisfied/Dissatisfied
- Factors : Sex, Race (White/Nonwhite),

  Age (<35, 35-44, >44)

  Region (Northeast, Mid-Atlantic, Southern, Midwest, Northwest, Southwest, Pacific)
- Explain Job satisfaction with all the main effects and the interactions.

Une approche exploratoire

(1) Régression PLS de

Y_1 = Logit(proportion of satisfied people)

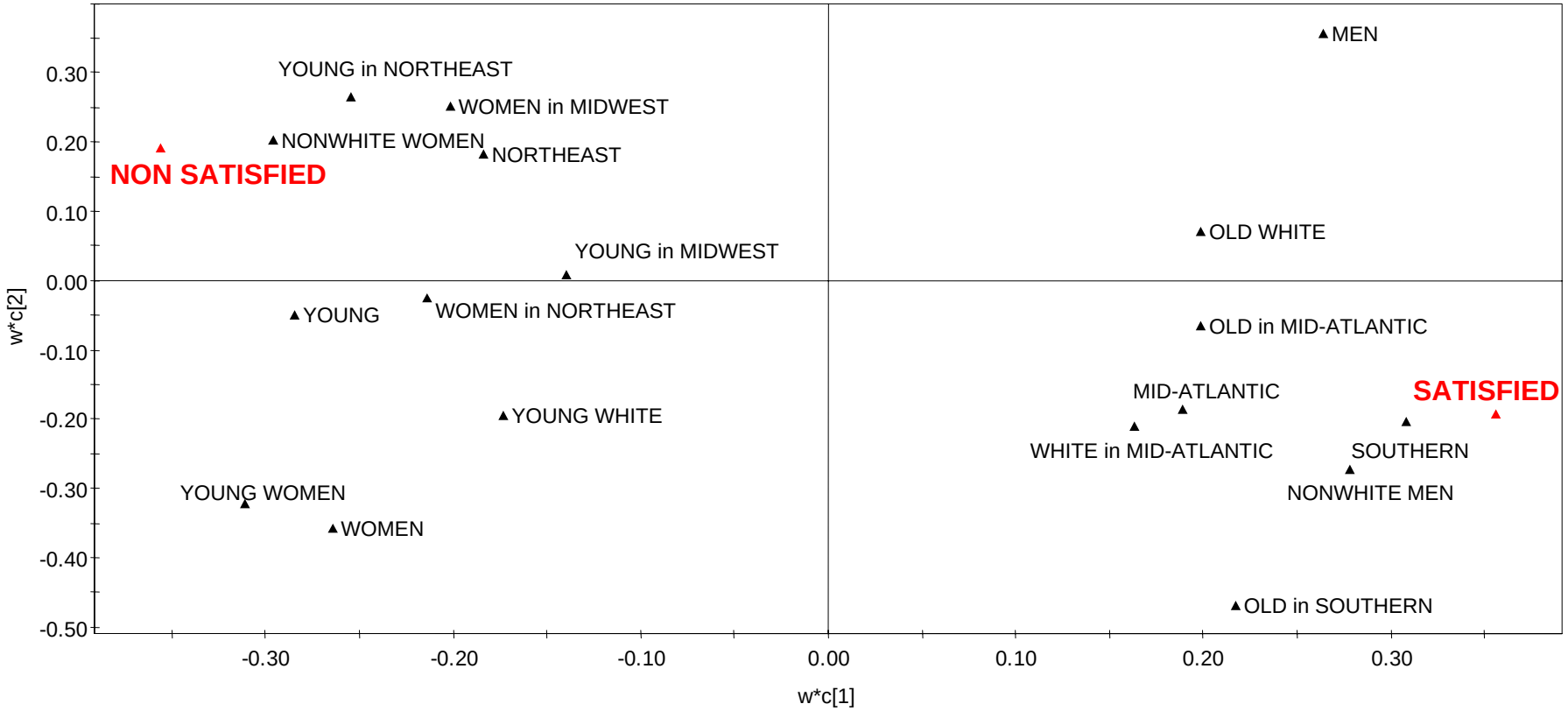
Y_2 = Logit(proportion of non satisfied people)

sur les 4 facteurs et toutes les interactions.

(2) Élimination itérative des termes à petits VIP, en vérifiant l'augmentation du Q^2(cum)

(3) Carte des variables finalement retenues

Résultat de la Régression PLS sur les logits



Y_1 = Logit (Proportion of Satisfied)

Y_2 = Logit (Proportion of Non Satisfied)

X = Explanatory variables kept after elimination of small VIP terms

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), Umetrics, 1999.
- H. Martens & M. Martens : Multivariate Analysis of Quality, Wiley, 2000
- H. Martens & T. Næs : Multivariate calibration, Wiley, 1989
- SIMCA 12.0 : PLS Software, S. WOLD, UMETRI (Sweden), distribué par SIGMA PLUS
- M. Tenenhaus : La régression PLS, Editions Technip, 1998

Approche PLS (PLS Path modelling)

- J.-B. Lohmöller : Latent variable path modeling with partial least squares, Physica-Verlag, 1989
- LVPLS 1.8 : Software for Latent variables path analysis with partial least-squares estimation, J.-B. Lohmöller, 1989
- M. Tenenhaus : L'approche PLS, R.S.A., 47 (2), 5-40, 1999