REVUE DE STATISTIQUE APPLIQUÉE

M. TENENHAUS
J.-P. GAUCHI
C. MÉNARDO

Régression PLS et applications

Revue de statistique appliquée, tome 43, n° 1 (1995), p. 7-63
<http://www.numdam.org/item?id=RSA_1995__43_1_7_0>

Rev. Statistique Appliquée, 1995, **XLIII** (1), 7-63

RÉGRESSION PLS ET APPLICATIONS

M. Tenenhaus (1), J.-P. Gauchi (2), C. Ménardo (3)

(1) Groupe HEC, Jouy-en-Josas
(2) Rhône-Poulenc, CRA, Aubervilliers
(3) Rhône-Poulenc, CRIT, Décines Charpieu

RÉSUMÉ

La régression PLS permet de relier un ensemble de variables dépendantes $Y = \{Y_1, \ldots, Y_P\}$ à un ensemble de variables indépendantes $X = \{X_1, \ldots, X_M\}$ lorsque le nombre de variables indépendantes et/ou dépendantes est élevé. La régression PLS consiste à effectuer une analyse en composantes principales de l'ensemble de variables X sous la contrainte que les (pseudo-) composantes principales des X_j soient aussi «explicatives» que possible de l'ensemble de variables Y. Il est alors possible de prédire les Y_k à partir des X_j en séparant mieux le signal (ce qui est synthétique, structuré, commun aux données) du bruit (ce qui est plus spécifique à l'échantillon de données étudié). Nous présentons dans cet article la régression PLS d'un point de vue «analyse des données», puis exposons deux applications industrielles, l'une réalisée au Centre de Recherches d'Aubervilliers et l'autre au Centre d'Industrialisation de Décines de Rhône-Poulenc. Enfin nous montrons comment utiliser la régression PLS usuelle lorsque les données sont qualitatives.

Mots-clés : *analyse discriminante PLS, régression PLS, régression sur composantes principales, régression qualitative PLS.*

SUMMARY

Partial Least-Squares regression makes it possible to relate a set of dependent variables $Y = \{Y_1, \ldots, Y_P\}$ to a set of independent variables $X = \{X_1, \ldots, X_M\}$, when there is a large number of independent and/or dependent variables. Partial Least-Squares regression consists in carrying out a principal component analysis of the set of variables X subject to the constraint that the (pseudo-) principal components of the X_j are as "explanatory" as possible of the set of variables Y. It is then possible to predict the Y_k from the X_j by better separating the signal (which is synthetic, structured, common to the data) from the noise (which is more specific to the sample under study). We present in this paper PLS regression with a "data analysis" point-of-view, and describe two industrial applications, one carried out at the Centre de Recherches d'Aubervilliers, and the other at the Centre d'Industrialisation de Décines, both at Rhône-Poulenc. At least we show how to use PLS regression when the data are categorical.

Keywords : *PLS discriminant analysis, PLS qualitative regression, PLS regression, regression on principal components.*

Introduction

La régression PLS (Partial Least Squares Regression) permet de relier un ensemble de variables dépendantes $Y = \{Y_1, \ldots, Y_P\}$ à un ensemble de variables indépendantes $X = \{X_1, \ldots, X_M\}$ lorsque le nombre de variables indépendantes et/ou dépendantes est élevé. La régression PLS consiste à effectuer une analyse en composantes principales de l'ensemble de variables X sous la contrainte que les (pseudo-)composantes principales des X_j soient aussi «explicatives» que possible de l'ensemble de variables Y. Il est alors possible de prédire les Y_k à partir des X_j en séparant mieux le signal (ce qui est synthétique, structuré, commun aux données) du bruit (ce qui est plus spécifique à l'échantillon de données étudié). Après avoir détaillé la théorie, nous illustrons l'utilisation de la méthode de régression PLS à l'aide de deux exemples tirés de la littérature et traités à l'aide du logiciel SIMCA. Dans une deuxième partie nous présentons deux applications industrielles, l'une réalisée au Centre de Recherches d'Aubervilliers et l'autre au Centre d'Industrialisation de Décines de Rhône-Poulenc, et montrons comment utiliser la régression PLS usuelle lorsque les données sont qualitatives. Ce travail s'appuie sur les références suivantes : Geladi et Kowalski (1986), Glen, Dunn III et Scott (1989), Höskuldsson (1988), Kvalheim (1988), Martens et Naes (1989) et Wold (1989). On trouvera une application détaillée de la régression PLS en chimie de formulation dans Gauchi (1995).

1. La régression PLS

1.1. Données et notations

Les notations de la régression PLS sont relativement bien standardisées dans la littérature. Les voici :

N = Nombre d'individus

M = Nombre de variables indépendantes X_j

P = Nombre de variables dépendantes Y_k

A = Nombre de composantes retenues

X = Matrice des données ($N \times M$) pour les variables indépendantes

Y = Matrice des données ($N \times P$) pour les variables dépendantes

E_0 = Matrice des variables X_j centrées-réduites

F_0 = Matrice des variables Y_k centrées-réduites

E_h = Matrice des résidus de la décomposition de E_0 en utilisant h composantes

E_{hj} = j-ième colonne de E_h

F_h = Matrice des résidus de la décomposition de F_0 en utilisant h composantes

F_{hk} = k-ième colonne de F_h.

1.2. La méthode

Il s'agit d'étudier les liaisons entre les variables dépendantes Y_k et les variables indépendantes X_j en prenant en compte les relations internes entre les variables de chaque groupe.

L'algorithme de la régression PLS est itératif. Nous allons en décrire le principe, puis étudier plus en détail les propriétés de la solution.

L'algorithme de la régression PLS

Etape 0 On part des tableaux E_0 et F_0.

Etape 1 On construit une combinaison linéaire u_1 des colonnes de F_0 et une combinaison linéaire t_1 des colonnes de E_0 maximisant cov (u_1, t_1) = cor $(u_1, t_1) \cdot \sqrt{\text{var}\,(u_1) \cdot \text{var}\,(t_1)}$. On obtient donc deux variables u_1 et t_1 aussi corrélées que possible et résumant au mieux les tableaux F_0 et E_0. On construit ensuite les régressions :

$$E_0 = t_1 p_1' + E_1$$
$$F_0 = t_1 r_1' + F_1$$

Etape 2 On reprend l'étape 1 en remplaçant les tableaux E_0 et F_0 par E_1 et F_1. On obtient donc deux nouvelles composantes : u_2, combinaison linéaire des colonnes de F_1, t_2, combinaison linéaire des colonnes de E_1. D'où les décompositions obtenues par régression :

$$E_0 = t_1 p_1' + t_2 p_2' + E_2$$
$$F_0 = t_1 r_1' + t_2 r_2' + F_2$$

On itère la procédure jusqu'à ce que les composantes $t_1, \ldots, t_A$ expliquent suffisamment F_0. Les composantes t_h sont des combinaisons linéaires des colonnes de E_0, et non corrélées entre elles. De la décomposition $F_0 = t_1 r_1' + \ldots + t_h r_h' + F_h$ on peut donc déduire les équations de régression PLS :

$$Y_k = \beta_{k0} + \beta_{k1} X_1 + \ldots + \beta_{kM} X_M + F_{hk}$$

1.3. Recherche et propriétés des composantes PLS $t_1, \ldots, t_A$ et $u_1, \ldots, u_A$

Nous allons décrire dans cette section la recherche des deux premières composantes PLS et donner ensuite des résultats généraux.

1.3.1. Recherche de t_1 et u_1

On recherche une composante $t_1 = E_0 w_1$ et une composante $u_1 = F_0 c_1$, avec $||w_1|| = ||c_1|| = 1$, telles que la covariance entre t_1 et u_1 soit maximum. On adopte ici la norme usuelle. Comme cov $(t_1, u_1) = \sqrt{\text{var}\,(t_1)\ \text{var}\,(u_1)}\ \text{cor}\,(t_1, u_1)$ on essaie ainsi de maximiser simultanément la variance expliquée par t_1, la variance expliquée par u_1 et la corrélation entre ces deux composantes. On cherche donc des vecteurs normés w_1 et c_1 maximisant $< t_1, u_1 >= ||t_1|| \cdot ||u_1|| \cdot \text{cor}\,(t_1, u_1)$. Utilisons la méthode des multiplicateurs de Lagrange. On pose

$$s = w_1' E_0' F_0 c_1 - \lambda_1 (w_1' w_1 - 1) - \lambda_2 (c_1' c_1 - 1)$$

On annule les dérivées partielles :

$$\frac{\partial s}{\partial \lambda_1} = -(w_1' w_1 - 1) = 0 \tag{1}$$

$$\frac{\partial s}{\partial \lambda_2} = -(c_1' c_1 - 1) = 0 \tag{2}$$

$$\frac{\partial s}{\partial w_1} = E_0' F_0 c_1 - 2\lambda_1 w_1 = 0 \tag{3}$$

$$\frac{\partial s}{\partial c_1} = F_0' E_0 w_1 - 2\lambda_2 c_1 = 0 \tag{4}$$

Des égalités (1) à (4) on déduit :

$$2\lambda_1 = 2\lambda_2 = \theta_1 = w_1' E_0' F_0 c_1 = < t_1, u_1 >$$

D'où les relations :

$$E_0' F_0 c_1 = \theta_1 w_1 \tag{5}$$

$$F_0' E_0 w_1 = \theta_1 c_1 \tag{6}$$

$$E_0' F_0 F_0' E_0 w_1 = \theta_1^2 w_1 \tag{7}$$

$$F_0' E_0 E_0' F_0 c_1 = \theta_1^2 c_1 \tag{8}$$

Ainsi w_1 est vecteur propre de $E_0' F_0 F_0' E_0$ associé à la plus grande valeur propre θ_1^2, et c_1 est vecteur propre de $F_0' E_0 E_0' F_0$ associé à la plus grande valeur propre θ_1^2.

On effectue ensuite deux régressions : de E_0 sur t_1, puis de F_0 sur t_1 :

$$E_0 = t_1 p_1' + E_1 \tag{9}$$

où

$$p_1 = E_0' t_1 / ||t_1||^2 \tag{10}$$

est le vecteur des coefficients de régression des E_{0j} sur t_1.

$$F_0 = t_1 r_1' + F_1 \tag{11}$$

où

$$r_1 = F_0' t_1 / ||t_1||^2 \tag{12}$$

est le vecteur des coefficients de régression des F_{0k} sur t_1.

On a les propriétés suivantes :

(a) $p_1' w_1 = 1$

(b) $r_1 = b_1 c_1$ où b_1 est le coefficient de régression de u_1 sur t_1.

Les vecteurs r_1 et c_1 sont donc colinéaires.

Montrons (a) : $p_1' w_1 = t_1' E_0 w_1 / ||t_1||^2 = t_1' t_1 / ||t_1||^2 = 1.$

Montrons (b) :
$$\begin{aligned} r_1 &= F_0' t_1 / ||t_1||^2 = F_0' E_0 w_1 / ||t_1||^2 \\ &= \theta_1 c_1 / ||t_1||^2 = b_1 c_1 \end{aligned}$$

où $b_1 = \theta_1 / ||t_1||^2 = u_1' t_1 / ||t_1||^2$ est bien le coefficient de régression de u_1 sur t_1.

La recherche de t_1 et u_1 est visualisée dans la figure 1, et les régressions de E_0 et F_0 sur t_1 dans la figure 2.

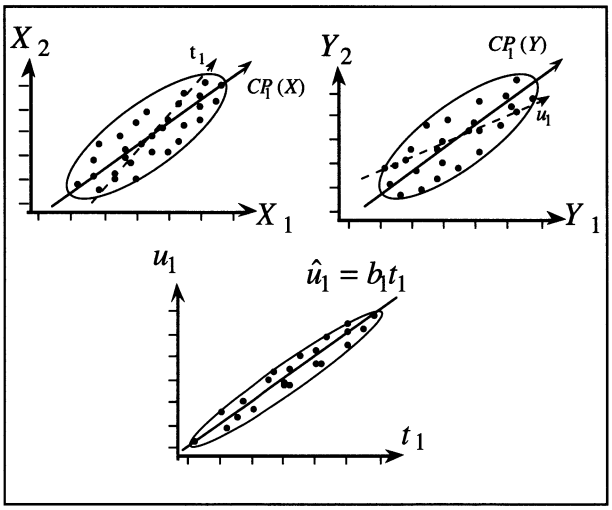


FIGURE 1
Visualisation de la recherche des composantes t_1 et u_1

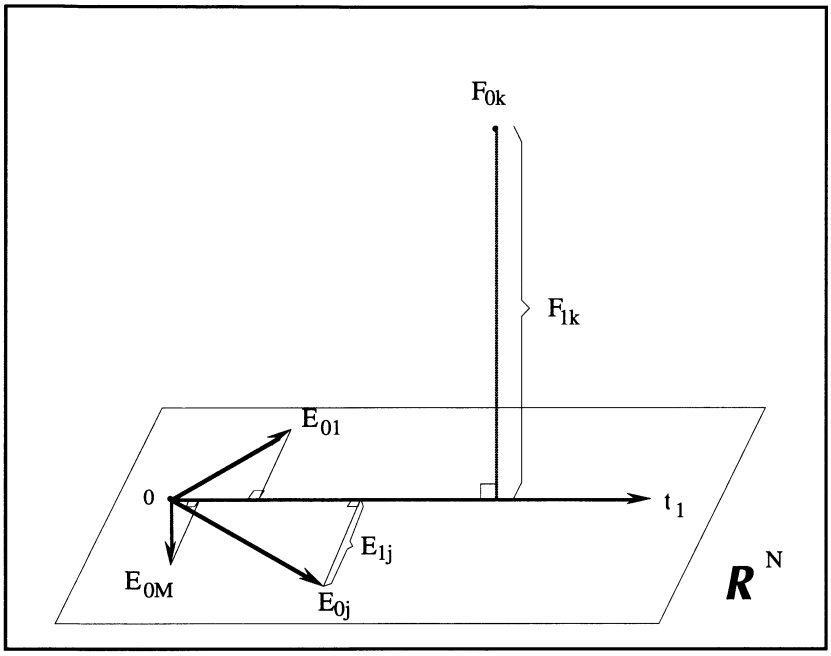


FIGURE 2
Régressions de E_0 et F_0 sur t_1

1.3.2. Recherche de t_2 et u_2

On remplace dans la section précédente E_0 et F_0 par les matrices résiduelles E_1 et F_1. D'où :

$$
\begin{aligned}
t_2 &= E_1 w_2 \\
u_2 &= F_1 c_2 \\
\theta_2 &= u_2' t_2 \\
p_2 &= E_1' t_2 / ||t_2||^2 \\
r_2 &= F_1' t_2 / ||t_2||^2 = \theta_2 c_2 / ||t_2||^2 = b_2 c_2
\end{aligned}
$$

où $b_2 = u_2' t_2 / ||t_2||^2$ est le coefficient de régression de u_2 sur t_2.

$$
\begin{aligned}
E_1 &= t_2 p_2' + E_2 \\
F_1 &= t_2 r_2' + F_2 \\
E_0 &= t_1 p_1' + t_2 p_2' + E_2 \\
F_0 &= t_1 r_1' + t_2 r_2' + F_2
\end{aligned}
$$

Et ainsi de suite pour les autres étapes.

1.3.3. Résultats généraux

Nous allons présenter dans cette section les propriétés des composantes PLS. Ces résultats sont nécessaires à une compréhension approfondie de la méthode.

On obtient les résultats suivants :

A l'étape h les équations (5) à (8) deviennent

$$E'_{h-1}F_{h-1}c_h = \theta_h w_h \tag{13}$$
$$F'_{h-1}E_{h-1}w_h = \theta_h c_h \tag{14}$$
$$E'_{h-1}F_{h-1}F'_{h-1}E_{h-1}w_h = \theta_h^2 w_h \tag{15}$$
$$F'_{h-1}E_{h-1}E'_{h-1}F_{h-1}c_h = \theta_h^2 c_h \tag{16}$$

Les équations (13) et (16) ou (14) et (15) permettent d'obtenir w_h et c_h et les composantes PLS $u_h = F_{h-1}c_h$ et $t_h = E_{h-1}w_h$, avec $\theta_h = u'_h t_h$.

Les équations (9) à (12) deviennent

$$E_{h-1} = t_h p'_h + E_h \tag{17}$$

où

$$p_h = E'_{h-1}t_h/||t_h||^2 \tag{18}$$
$$F_{h-1} = t_h r'_h + F_h \tag{19}$$

où

$$r_h = F'_{h-1}t_h/||t_h||^2 \tag{20}$$
$$= b_h c_h \tag{21}$$

avec b_h coefficient de régression de u_h sur t_h.

On a aussi :

$$w'_h p_h = 1 \tag{22}$$

et

$$t'_h E_h = 0, \tag{23}$$

propriété de la régression.

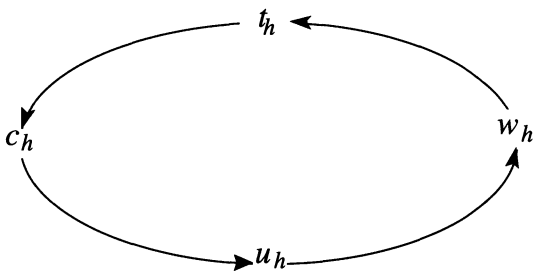
Relations cycliques

Partant de w_h on obtient la suite de relations suivantes :

$$t_h = E_{h-1}w_h$$
$$c_h = \frac{1}{\theta_h}F'_{h-1}t_h \tag{24}$$
$$u_h = F_{h-1}c_h$$
$$w_h = \frac{1}{\theta_h}E'_{h-1}u_h \tag{25}$$

La relation (24) provient de (14) et (25) de (13).

On peut schématiser



Relations d'orthogonalité

Il y a de nombreuses relations d'orthogonalité au niveau de l'analyse des variables X_j. Ceci n'est plus vrai au niveau de l'analyse des Y_k.

On a les relations d'orthogonalité suivantes :

$$\begin{aligned}
&\text{(a)} \quad w_h' w_l = 0, \quad l > h \\
&\text{(b)} \quad t_h' t_l = 0, \quad l > h \\
&\text{(c)} \quad w_h' p_l = 0, \quad l > h \\
&\text{(d)} \quad w_h' E_l' = 0, \quad l \geqslant h \\
&\text{(e)} \quad t_h' E_l = 0, \quad l \geqslant h
\end{aligned}$$

Preuve :

1) Démonstration de $t_h' t_l = 0, \quad l > h$

La preuve se fait par récurrence.

– $t_1' t_2 = t_1' E_1 w_2 = 0$, puisque $t_1' E_1 = 0$.

– Supposons $t_1, \ldots, t_h$ orthogonaux, alors : $t_1, \ldots, t_{h+1}$ sont orthogonaux.

Montrons que t_{h+1} est orthogonal aux vecteurs $t_1, \ldots, t_h$:

$$\begin{aligned}
t_h' t_{h+1} &= t_h' E_h w_{h+1} = 0, \text{ puisque } t_h' E_h = 0. \\
t_{h-1}' t_{h+1} &= t_{h-1}' E_h w_{h+1} \\
&= t_{h-1}' [E_{h-1} - t_h p_h'] w_{h+1} \\
&= [t_{h-1}' E_{h-1} - t_{h-1}' t_h p_h'] w_{h+1} \\
&= 0
\end{aligned}$$

puisque $t'_{h-1}E_{h-1} = 0 \quad \text{et} \quad t'_{h-1}t_h = 0.$

$$\begin{aligned} t'_{h-2}t_{h+1} &= t'_{h-2}E_h w_{h+1} \\ &= t'_{h-2}[E_{h-2} - t_{h-1}p'_{h-1} - t_h p'_h]w_{h+1} \\ &= 0 \end{aligned}$$

et ainsi de suite, d'où le résultat.

2) Démonstration de $w'_h E'_l = 0, \quad l \geqslant h$

On a :

$$\begin{aligned} w'_h E'_h &= w'_h[E_{h-1} - t_h p'_h]' \\ &= t'_h - w'_h p_h t'_h = 0 \end{aligned}$$

puisque $w'_h p_h = 1$.

Montrons maintenant que $w'_h E'_l = 0, \quad l \geqslant h$, implique $w'_h E'_{l+1} = 0$. On a :

$$\begin{aligned} w'_h E'_{l+1} &= w'_h[E_l - t_{l+1}p'_{l+1}]' \\ &= w'_h E'_l - \frac{w'_h E'_l t_{l+1}}{||t_{l+1}||^2} \cdot t'_{l+1} \\ &= 0, \end{aligned}$$

d'où le résultat.

3) Démonstration de $w'_h p_l = 0, \quad l > h$

On a :

$$\begin{aligned} w'_h p_l &= w'_h E'_{l-1} t_l / ||t_l||^2 \\ &= 0, \quad \text{pour } l > h \end{aligned}$$

puisque $w'_h E'_{l-1} = 0$, pour $l - 1 \geqslant h$, soit $l > h$.

4) Démonstration de $w'_h w_l = 0, \quad l > h$

On a :

$$\begin{aligned} w'_h w_l &= w'_h \cdot \frac{1}{\theta_l} E'_{l-1} u_l \\ &= 0, \quad l > h \end{aligned}$$

puisque $w'_h E'_{l-1} = 0$ pour $l - 1 \geqslant h$, soit $l > h$.

5) Démonstration de $t'_h E_l = 0, \quad l \geqslant h$

L'égalité est vraie pour $l = h$.

Si $l > h$ on a :

$$\begin{aligned} t'_h E_l &= t'_h(E_{l-1} - t_l p'_l) \\ &= t'_h E_{l-1} \\ &= t'_h(E_{l-2} - t_{l-1} p'_{l-1}) \\ &= t'_h E_{l-2} = \ldots = t'_h E_h = 0. \end{aligned}$$

Formules de décomposition

Les matrices E_0 et F_0 peuvent se décomposer par régression sur les composantes $t_1, \ldots, t_A$, où A est le rang de E_0 :

$$E_0 = t_1 p'_1 + t_2 p'_2 + \ldots + t_A p'_A \tag{26}$$
$$F_0 = t_1 r'_1 + t_2 r'_2 + \ldots + t_A r'_A + F_A \tag{27}$$

Les vecteurs t_h étant orthogonaux entre eux on retrouve bien les mêmes coefficients de régression dans les équations (26) et (27) que dans la méthode itérative décrite plus haut.

La matrice des résidus E_h à l'étape h peut s'exprimer en fonction de E_0 :

$$E_h = E_0(I - w_1 p'_1)(I - w_2 p'_2) \ldots (I - w_h p'_h) \tag{28}$$

Montrons ce résultat.

On a $E_1 = E_0 - t_1 p'_1 = E_0 - E_0 w_1 p'_1 = E_0(I - w_1 p'_1)$

Supposons la décomposition (28) vraie pour $h = j$. Montrons qu'elle est vraie pour $h = j + 1$:

$$\begin{aligned} E_{j+1} &= E_j - t_{j+1} p'_{j+1} = E_j - E_j w_{j+1} p'_{j+1} \\ &= E_j(I - w_{j+1} p'_{j+1}) \\ &= E_0 \prod_{h=1}^{j} (I - w_h p'_h)(I - w_{j+1} p'_{j+1}) \end{aligned}$$

d'où le résultat.

Ces décompositions entraînent trois résultats importants au niveau de l'interprétation. Notons $||E_0||^2$ la somme des carrés des différents éléments de la matrice E_0. On a alors :

$$||E_0||^2 = ||t_1||^2 \cdot ||p_1||^2 + ||t_2||^2 \cdot ||p_2||^2 + \ldots + ||t_A||^2 \cdot ||p_A||^2 \tag{29}$$

Cette formule permet de mesurer le pouvoir de chaque t_h (ou des h premiers) pour résumer E_0. On utilise

$$\frac{||t_h||^2 \cdot ||p_h||^2}{||E_0||^2} \quad \text{ou} \quad \frac{\sum_{l \leqslant h} ||t_h||^2 \cdot ||p_h||^2}{||E_0||^2}$$

Montrons (29) :

Les $t_1, \ldots, t_A$ étant orthogonaux on obtient pour chaque colonne E_{0j}, en utilisant (26) :

$$||E_{0j}||^2 = \sum_{h=1}^{A} ||t_h||^2 p_{hj}^2$$

et par conséquent

$$||E_0||^2 = \sum_{h=1}^{A} ||t_h||^2 \cdot ||p_h||^2.$$

On aura de même la décomposition :

$$||F_0||^2 = ||t_1||^2 \cdot ||r_1||^2 + \ldots + ||t_A||^2 \cdot ||r_A||^2 + ||F_A||^2 \qquad (30)$$

On mesure le pouvoir de chaque t_h (ou des h premiers) pour expliquer F_0 à l'aide de

$$\frac{||t_h||^2 \cdot ||r_h||^2}{||F_0||^2} \quad \text{ou} \quad \frac{\sum_{l \leqslant h} ||t_h||^2 \cdot ||r_h||^2}{||F_0||^2}$$

Enfin il est important de remarquer que la composante t_h est combinaison linéaire des colonnes de E_0 :

$$\begin{aligned} t_h &= E_{h-1} w_h \\ &= E_0 \prod_{l=1}^{h-1} (I - w_l p_l') w_h = E_0 \widetilde{w}_h \end{aligned}$$

Le vecteur $\widetilde{w}_h = \prod_{l=1}^{h-l} (I - w_l p_l') w_h$ des coefficients des variables E_{0j} n'est cependant pas normé à 1.

Les composantes PLS $t_1, \ldots, t_h$ sont donc des combinaisons linéaires des colonnes de E_0, non corrélées entre elles, résumant au mieux E_0 tout en expliquant

autant que possible F_0. Ces composantes PLS sont donc analogues à des composantes principales des $X_1, \ldots, X_M$ expliquant au mieux les variables $Y_1, \ldots, Y_P$.

Les composantes PLS $u_1, \ldots, u_h$ n'ont pas cette propriété. La composante u_1 est bien combinaison linéaire des colonnes de F_0, mais ce n'est plus le cas des autres composantes. Les composantes $u_2, \ldots, u_h$ sont des variables intermédiaires plus difficiles à interpréter.

Les graphiques

Le plan (t_1, t_2) permet de visualiser les individus dans un plan résumant au mieux les X_j, tout en étant orienté vers l'explication des Y_k.

Le plan (t_1, u_1) permet de visualiser la liaison entre les Y_k et les X_j détectée par ces premières composantes PLS.

Pour interpréter les composantes PLS t_h en fonction des X_j et des Y_k il est naturel de calculer les corrélations entre X_j et t_h, puis entre Y_k et t_h. On obtient :

$$\text{cor}\,(X_j, t_h) = \frac{\frac{1}{N-1} t_h' E_{0j}}{\sqrt{\frac{1}{N-1} t_h' t_h}} = \frac{\frac{1}{N-1} t_h' t_h p_{hj}}{\sqrt{\frac{1}{N-1} t_h' t_h}} = \sqrt{\text{Var}\,(t_h)} \cdot p_{hj}$$

Bien que la division par $N-1$ plutôt que N ne s'impose pas, nous l'avons utilisée pour être cohérent avec les sorties du logiciel SIMCA.

Notant λ_h la variance de t_h on obtient :

$$\text{cor}\,(X_j, t_h) = \sqrt{\lambda_h} \cdot p_{hj}$$

Et de même

$$\text{cor}\,(Y_k, t_h) = \sqrt{\lambda_h} \cdot r_{hk}$$

Pour interpréter t_1 et t_2, on construit donc le cercle des corrélations en visualisant dans un plan les points

$$A_j = (\sqrt{\lambda_1} \cdot p_{1j}, \sqrt{\lambda_2} \cdot p_{2j})$$

et

$$B_k = (\sqrt{\lambda_1} \cdot r_{1k}, \sqrt{\lambda_2} \cdot r_{2k})$$

représentant les corrélations de X_j et Y_k avec t_1 et t_2.

Les produits scalaires $< A_j, A_{j'} >$, $< B_k, B_{k'} >$ et $< A_j, B_k >$ représentent des approximations à l'ordre 2 des cor $(X_j, X_{j'})$, cor $(Y_k, Y_{k'})$ et cor (X_j, Y_k) respectivement.

En effet :

$$\begin{aligned}\text{cor}\,(X_j, X_{j'}) &= \frac{1}{N-1} E'_{0j} E_{0j'} \\ &= \lambda_1 p_{1j} p_{1j'} + \lambda_2 p_{2j} p_{2j'} + \ldots + \lambda_A p_{Aj} p_{Aj'} \\ &\approx < A_j, A_{j'} >\end{aligned}$$

On démontre de même les autres approximations.

La norme de A_j représente la corrélation multiple entre X_j et (t_1, t_2) et celle de B_k la corrélation multiple entre Y_k et (t_1, t_2). On a en effet

$$R^2(X_j; t_1, t_2) = \text{cor}^2(X_j, t_1) + \text{cor}^2(X_j, t_2)$$

puisque les variables t_1 et t_2 sont non corrélées. D'où le résultat. Et de même pour Y_k.

Ainsi le cercle des corrélations indique les variables bien corrélées aux deux premières composantes PLS. Pour les variables bien expliquées par t_1 et t_2, le cercle des corrélations traduit aussi bien les corrélations internes à chaque groupe de variables que les corrélations inter-groupes.

1.4. Les équations de régression PLS

De la décomposition de F_0 sur $t_1, \ldots, t_h$ on déduit la régression PLS de chaque variable Y_k sur $X_1, \ldots, X_M$. On part de

$$\begin{aligned}F_{0k} = \frac{Y_k - \overline{y}_k}{s_{y_k}} &= \sum_{l=1}^{h} r_{lk} t_l + F_{hk} \\ &= \sum_{l=1}^{h} r_{lk} E_0 \widetilde{w}_l + F_{hk} \\ &= \sum_{l=1}^{h} r_{lk} \sum_{j=1}^{M} \widetilde{w}_{lj} \left(\frac{X_j - \overline{x}_j}{s_{x_j}} \right) + F_{hk}\end{aligned}$$

Posons $\beta_j = \sum_{l=1}^{h} r_{lk} \widetilde{w}_{lj}$.

On trouve dans les programmes de régression PLS trois types de résultats :

(1) Les variables Y_k et les X_j sont centrées-réduites.

D'où :

$$F_{0k} \approx \sum_{j=1}^{M} \beta_j E_{0j}$$

(2) On utilise les variables réduites $Y'_k = \dfrac{Y_k}{s_{y_k}}$ et $X'_j = \dfrac{X_j}{s_{x_j}}$.

D'où :

$$Y'_k \approx \beta_0 + \sum_{j=1}^{M} \beta_j X'_j$$

où $\beta_0 = \dfrac{\overline{y}_k}{s_{y_k}} - \sum_{j=1}^{M} \beta_j \overline{x}_j / s_{x_j}$

Les coefficients de régression β_j sont comparables et donc interprétables au niveau de l'explication de Y_k par les X_j.

(3) On utilise les variables d'origine.

D'où :

$$Y_k \approx \beta'_0 + \sum_{j=1}^{M} \beta'_j X_j$$

où $\beta'_0 = \beta_0 s_{y_k}$ et $\beta'_j = \beta_j \dfrac{s_{y_k}}{s_{x_j}}$.

1.5. Le cas particulier d'une seule variable dépendante

La régression PLS est particulièrement utile lorsqu'on cherche à expliquer une variable Y à l'aide de plusieurs variables X_j très corrélées entre elles, ou même linéairement dépendantes, et qu'on ne souhaite pas abandonner des variables explicatives. Les formules des sections précédentes se simplifient. Nous allons donc reprendre les étapes de la méthode de régression PLS pour ce cas particulier.

Etape 1

La matrice F_0 se réduit à une colonne $F_0 = Y^* = \dfrac{Y - \overline{y}}{s_y}$. Par conséquent $c_1 = 1$ et $u_1 = F_0$. On recherche une composante $t_1 = E_0 w_1$, avec $||w_1|| = 1$, telle que la covariance entre t_1 et u_1 soit maximum. Comme cov $(t_1, u_1) = \sqrt{\text{Var}\,(t_1)} \times \text{cor}\,(t_1, u_1)$ on essaie de maximiser simultanément la variance expliquée par t_1 et la corrélation entre Y et t_1. On réalise ainsi un compromis entre la régression multiple de Y sur $X_1, \ldots, X_M$ et l'analyse en composantes principales des $X_1, \ldots, X_M$. L'équation (5) permet d'obtenir directement w_1 puisque $c_1 = 1$ et que $||w_1|| = 1$:

$$w_1 = \frac{E'_0 F_0}{||E'_0 F_0||}$$

De $E_0'F_0 = (N-1)[\text{cor}\,(X_1,Y),\ldots,\text{cor}\,(X_M,Y)]'$ on déduit :

$$w_1 = \frac{1}{\sqrt{\sum_{j=1}^{M}\text{cor}^2(X_j,Y)}} \begin{bmatrix} \text{cor}\,(X_1,Y) \\ \vdots \\ \text{cor}\,(X_M,Y) \end{bmatrix}$$

et

$$\begin{aligned} \theta_1 =< t_1,u_1 > &= w_1'E_0'F_0 \\ &= ||w_1||^2 \times ||E_0'F_0|| \\ &= ||E_0'F_0|| \\ &= (N-1)\sqrt{\sum_{j=1}^{M}\text{cor}^2(X_j,Y)} \end{aligned}$$

Par conséquent

$$\text{cov}\,(t_1,u_1) = \sqrt{\sum_{j=1}^{M}\text{cor}^2(X_j,Y)}$$

On obtient une composante t_1 qui s'écrit

$$\begin{aligned} t_1 &= E_0w_1 \\ &= \frac{1}{\sqrt{\sum_{j=1}^{M}\text{cor}^2(X_j,Y)}} \sum_{j=1}^{M}\text{cor}\,(X_j,Y)E_{0j} \end{aligned}$$

dont l'interprétation est tout-à-fait naturelle. On effectue ensuite les régressions de E_{0j} sur t_1, puis de F_0 sur t_1 :

$$\begin{aligned} E_{0j} &= p_{1j}t_1 + E_{1j} \\ F_0 &= r_1t_1 + F_1 \end{aligned}$$

où $p_{1j} = E_{0j}'t_1/||t_1||^2$ est le coefficient de régression de E_{0j} sur t_1 et $r_1 = F_0't_1/||t_1||^2$ le coefficient de régression de F_0 sur t_1. On pose :

$$\begin{aligned} E_1 &= [E_{11},\ldots,E_{1M}] \\ p_1 &= [p_{11},\ldots,p_{1M}]' \end{aligned}$$

D'où les décompositions :

$$E_0 = t_1 p_1' + E_1$$
$$F_0 = r_1 t_1 + F_1$$

Etape 2

On remplace dans la première étape E_0 et F_0 par la matrice résiduelle E_1 et le vecteur résiduel F_1. On obtient $c_2 = 1$, $u_2 = F_1$ et $w_2 = \dfrac{E_1' F_1}{||E_1' F_1||}$.

De $E_1' F_1 = (N-1)[\text{cov}\,(E_{11}, F_1), \ldots, \text{cov}\,(E_{1M}, F_1)]'$ on déduit

$$w_2 = \frac{1}{\sqrt{\sum_{j=1}^{M} \text{cov}^2(E_{1j}, F_1)}} \begin{bmatrix} \text{cov}\,(E_{11}, F_1) \\ \vdots \\ \text{cov}\,(E_{1M}, F_1) \end{bmatrix}$$

et

$$t_2 = E_1 w_2 = E_0 (I - w_1 p_1') w_2$$

combinaison linéaire des colonnes de E_0.

On pose :

$$p_2 = \frac{E_1' t_2}{||t_2||^2}$$
$$r_2 = \frac{F_1' t_2}{||t_2||^2} = \text{coefficient de régression de } F_1 \text{ sur } t_2.$$

D'où les décompositions :

$$E_1 = t_2 p_2' + E_2$$
$$F_1 = r_2 t_2 + F_2$$

et ainsi de suite pour les autres étapes.

Résultats généraux

On obtient les décompositions suivantes à l'ordre h :

$$E_0 = t_1 p_1' + t_2 p_2' + \ldots + t_h p_h' + E_h$$
$$F_0 = Y^* = r_1 t_1 + r_2 t_2 + \ldots + r_h t_h + F_h$$

Chaque composante t_h étant combinaison linéaire des $X_j^* = E_{0j}$, on peut écrire l'équation de régression PLS

$$Y^* = \beta_1 X_1^* + \ldots + \beta_M X_M^* + F_h$$

1.6. Exemple 1

On trouve dans Kettaneh-Wold [1992] l'exemple suivant tiré de Cornell [1990] :

L'indice d'octane moteur de douze différents mélanges a été enregistré afin de déterminer l'influence des sept composants suivants :

Distillation directe	$0 \leqslant x_1 \leqslant 0.21$
Reformat	$0 \leqslant x_2 \leqslant 0.62$
Naphta de craquage thermique	$0 \leqslant x_3 \leqslant 0.12$
Naphta de craquage catalytique	$0 \leqslant x_4 \leqslant 0.62$
Polymère	$0 \leqslant x_5 \leqslant 0.12$
Alkylat	$0 \leqslant x_6 \leqslant 0.74$
Essence naturelle	$0 \leqslant x_7 \leqslant 0.08$

Les douze mélanges (Tableau 1) ont été choisis selon un plan d'expériences D-optimal, en sélectionnant une partie des sommets.

TABLEAU 1
Données de Cornell

x_j n°	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
n°1	0.00	0.23	0.00	0.00	0.00	0.74	0.03	98.7
n°2	0.00	0.10	0.00	0.00	0.12	0.74	0.04	97.8
n°3	0.00	0.00	0.00	0.10	0.12	0.74	0.04	96.6
n°4	0.00	0.49	0.00	0.00	0.12	0.37	0.02	92.0
n°5	0.00	0.00	0.00	0.62	0.12	0.18	0.08	86.6
n°6	0.00	0.62	0.00	0.00	0.00	0.37	0.01	91.2
n°7	0.17	0.27	0.10	0.38	0.00	0.00	0.08	81.9
n°8	0.17	0.19	0.10	0.38	0.02	0.06	0.08	83.1
n°9	0.17	0.21	0.10	0.38	0.00	0.06	0.08	82.4
n°10	0.17	0.15	0.10	0.38	0.02	0.10	0.08	83.2
n°11	0.21	0.36	0.12	0.25	0.00	0.00	0.06	81.4
n°12	0.00	0.00	0.00	0.55	0.00	0.37	0.08	88.1

On calcule la matrice des corrélations :

TABLEAU 2
Matrice des corrélations

	x_2	x_3	x_4	x_5	x_6	x_7	y
x_1	.10	**.999**	.37	-.55	**-.80**	**.60**	**-.84**
x_2		.10	-.54	-.29	-.19	**-.59**	-.07
x_3			.37	-.55	**-.80**	**.61**	**-.84**
x_4				-.21	**-.64**	**.92**	**-.71**
x_5					.46	-.27	.49
x_6						**-.66**	**.98**
x_7							**-.74**

Les corrélations les plus importantes sont en gras.

La régression PLS va nous permettre de relier la variable y à *l'ensemble* des variables explicatives $x_1, \ldots, x_7$.

1.6.1. Etude de la première dimension

De la formule

$$w_1 = \frac{1}{\sqrt{\sum_{j=1}^{M} \text{cor}^2(X_j, Y)}} \begin{bmatrix} \text{cor}\,(X_1, Y) \\ \vdots \\ \text{cor}\,(X_M, Y) \end{bmatrix}$$

on déduit :

$$w_1 = [-.437, -.037, -.437, -.369, .258, .514, -.387]'$$

et

$$t_1 = -.437E_{01} - .037E_{02} - .437E_{03} - .369E_{04} + .258E_{05} + .514E_{06} - .387E_{07},$$

où E_{0j} est la variable X_j centrée-réduite.

La composante t_1 calculée sur les douze expériences de l'exemple de Cornell est donnée dans le tableau 3.

Dans le logiciel SIMCA le calcul de $u_1 = F_0 c_1$ n'impose pas au vecteur c_1 d'être unitaire. La contrainte consiste à imposer au coefficient de régression de u_1 sur t_1 d'être égal à 1 :

$$\frac{u_1' t_1}{||t_1||^2} = \frac{c_1 F_0' t_1}{||t_1||^2} = 1$$

D'où on déduit que :

$$c_1 = [\text{coefficient de régression de } F_0 \text{ sur } t_1]^{-1} = \frac{1}{r_1} = \frac{1}{0.482}$$

On obtient donc :

$$u_1 = \frac{F_0}{r_1} = \frac{1}{r_1} \times \frac{y - \overline{y}}{s_y} = \frac{1}{0.482} \times \frac{y - 88.5833}{6.52127}$$

On trouve u_1 dans le tableau 3.

TABLEAU 3
Les composantes de la régression PLS

Obs.	t_1	u_1	t_2	u_2	t_3	u_3
1	2.051	3.218	.821	2.059	1.582	3.280
2	2.475	2.932	.649	.807	.109	.420
3	2.331	2.550	.927	.387	-.173	-1.431
4	2.037	1.087	-1.596	-1.677	-.502	-.216
5	-.068	-.631	-.218	-.993	-2.956	-2.055
6	1.614	.832	-1.423	-1.379	.971	.116
7	-2.204	-2.126	-.178	.138	.238	.838
8	-1.994	-1.744	.101	.440	.118	.899
9	-2.088	-1.967	.149	.213	.355	.170
10	-1.916	-1.713	.318	.359	.196	.107
11	-2.076	-2.285	-.461	-.369	1.031	.244
12	-.162	-.154	.910	.015	-.970	-2.371

On effectue ensuite les régressions des variables E_{0j} et F_0 sur t_1 :

$$E_{0j} = p_{1j} t_1 + E_{1j}, \qquad j = 1, \ldots, 7$$
$$F_0 = r_1 t_1 + F_1$$

On réalise donc huit régressions simples.

On obtient :

$$p_1 = \begin{bmatrix} -.454 \\ .0317 \\ -.454 \\ -.356 \\ .294 \\ .461 \\ -.413 \end{bmatrix} \qquad r_1 = 0.482$$

et les huit colonnes de résidus $E_{11}, \ldots, E_{17}, F_1$.

Pour évaluer la qualité de la première composante t_1 on calcule :

• *Part de variance de Y expliquée par t_1*

$$R^2(Y, t_1) = 0.924$$

• *Part de variance de X expliquée par t_1*

La décomposition $E_0 = t_1 p_1' + E_1$ donne $||E_0||^2 = ||t_1||^2 \times ||p_1||^2 + ||E_1||^2$, soit en divisant par $N - 1$, les variables E_{0j} étant centrées-réduites,

$$M = \frac{1}{N-1}||t_1||^2 \times ||p_1||^2 + \frac{1}{N-1}||E_1||^2$$

La part de variance de X expliquée par t_1 vaut donc :

$$\frac{||t_1||^2 \times ||p_1||^2}{(N-1)M} = \frac{43.72 \times 1.01}{11 \times 7} = 0.574$$

1.6.2. Etude de la deuxième dimension

On part des résidus E_1 et F_1 de la décomposition à une composante :

$$E_0 = t_1 p_1' + E_1$$
$$F_0 = r_1 t_1 + F_1$$

On obtient :

	1	2	3	4	5	6	7
cov (E_{1j}, F_1)	.03174	-.13153	.03261	-.02452	-.0701	.09992	.04933

et

$$\sum_{j=1}^{7} \text{cov}^2(E_{1j}, F_1) = 0.0373.$$

De la formule

$$w_2 = \frac{1}{\sqrt{\sum_{j=1}^{M} \text{cov}^2(E_{1j}, F_1)}} \begin{bmatrix} \text{cov}\,(E_{11}, F_1) \\ \vdots \\ \text{cov}\,(E_{1M}, F_1) \end{bmatrix}$$

on déduit :

$$w_2 = [.164, -.681, .169, -.127, -.363, .516, .255]'$$

et :

$$t_2 = .164E_{11} - .681E_{12} + .169E_{13} - .127E_{14} - .363E_{15} + .516E_{16} + .255E_{17}$$

On trouve la composante t_2 dans le tableau 3.

On décompose (régresse) E_1 et F_1 sur t_2 :

$$E_1 = t_2 p_2' + E_2$$
$$F_1 = t_2 r_2 + F_2$$

On obtient

$$p_2 = \begin{bmatrix} -0.042 \\ -1.003 \\ -0.039 \\ 0.278 \\ -0.045 \\ 0.440 \\ 0.477 \end{bmatrix} \qquad \text{et} \qquad r_2 = 0.273.$$

On calcule $u_2 = F_1/r_2$ de manière à avoir une régression de u_2 sur t_2 ayant un coefficient de régression égal à 1. La composante u_2 est donnée dans le tableau 3.

Qualité de la décomposition sur t_1, t_2

• Part de variance de Y expliquée par t_2

$$R^2(Y, t_2) = 0.053$$

• Part de variance de Y expliquée par t_1, t_2

$$\begin{aligned} R^2(Y; t_1, t_2) &= R^2(Y, t_1) + R^2(Y, t_2) \\ &= 0.924 + 0.053 = 0.977 \end{aligned}$$

• Part de variance de X expliquée par t_2

$$\frac{||t_2||^2 \times ||p_2||^2}{(N-1)M} = 0.153$$

• Part de variance de X expliquée par t_1, t_2

$$\frac{||t_1||^2 \times ||p_1||^2 + ||t_2||^2 \times ||p_2||^2}{(N-1)M} = 0.574 + 0.153 = 0.727.$$

1.6.3. Etude de la troisième dimension

On obtient comme précédemment les composantes t_3 et u_3 données dans le tableau 3.

Le vecteur w_3 vaut :

$$w_3 = [.290, .451, .291, -.567, -.445, .109, -.310]'$$

Qualité de la décomposition sur t_1, t_2, t_3

• Part de variance de Y expliquée par t_3

$$R^2(Y, t_3) = 0.014$$

• Part de variance de Y expliquée par t_1, t_2, t_3

$$R^2(Y; t_1, t_2, t_3) = 0.991$$

• Part de variance de X expliquée par t_3

$$\frac{||t_3||^2 \times ||p_3||^2}{(N-1)M} = 0.192$$

• Part de variance de X expliquée par t_1, t_2, t_3

$$\sum_{h=1}^{3} \frac{||t_h||^2 \times ||p_h||^2}{(N-1)M} = 0.727 + 0.192 = 0.919$$

Bien que l'apport de la troisième composante soit minime dans l'explication de Y, nous l'avons conservée pour retrouver les résultats de l'article de Kettaneh-Wold.

1.6.4. L'équation de régression PLS

On étudie la décomposition de F_0 sur t_1, t_2, t_3 :

$$F_0 = r_1 t_1 + r_2 t_2 + r_3 t_3 + F_3$$

En revenant aux variables de départ, il vient

$$y \approx 92.676 - 9.828x_1 - 6.96x_2 - 16.67x_3 - 8.422x_4 - 4.389x_5 + 10.16x_6 - 34.53x_7$$

On peut aussi écrire le modèle sur les variables réduites afin de mieux mettre en évidence l'influence marginale de chaque variable :

$$y' \approx 14.21 - .139x'_1 - .2087x'_2 - .1376x'_3 - .2932x'_4 - .0384x'_5 + .4564x'_6 - .1434x'_7$$

où $y' = y/s_y$ et $x'_j = x_j/s_{x_j}$ représentent les variables y et x_j réduites.

L'examen de la matrice des corrélations montre la cohérence de l'équation de régression sur les variables réduites.

(1) signes cohérents des coefficients de régression

(2) rôle négligeable de x_5 sur y

(3) forte influence de la variable x_6 sur y.

1.6.5. Les graphiques

Nous commentons dans cette section différents graphiques fournis par le logiciel SIMCA et le cercle des corrélations non fourni par SIMCA.

Le graphique (t_1, u_1) représenté dans la figure 3 illustre la forte corrélation entre Y et t_1.

Il est possible de prédire les trois classes de valeurs de y à l'aide des x_j :

$$\begin{array}{llll} 1,2,3,4,6 & \rightarrow & y \text{ dans } & [91.2 - 98.7] \\ 5,12 & \rightarrow & y \text{ dans } & [86.6 - 88.1] \\ 7,8,9,10,11 & \rightarrow & y \text{ dans } & [81.4 - 83.2] \end{array}$$

Le deuxième graphique (figure 4) isole les objets 4 et 6 qui sont effectivement en bas de la première classe au niveau du y : $y_4 = 92$ et $y_6 = 91.2$.

Le graphique de la figure 5 est moins intéressant : toutefois, 5 apparaît séparé de 12.

Le graphique (t_1, t_2) de la figure 6 est l'équivalent d'un premier plan principal orienté vers l'explication de y. On retrouve clairement les différentes classes d'objets identifiées ci dessus.

Enfin le cercle des corrélations de la figure 7 est construit à partir du tableau 4.

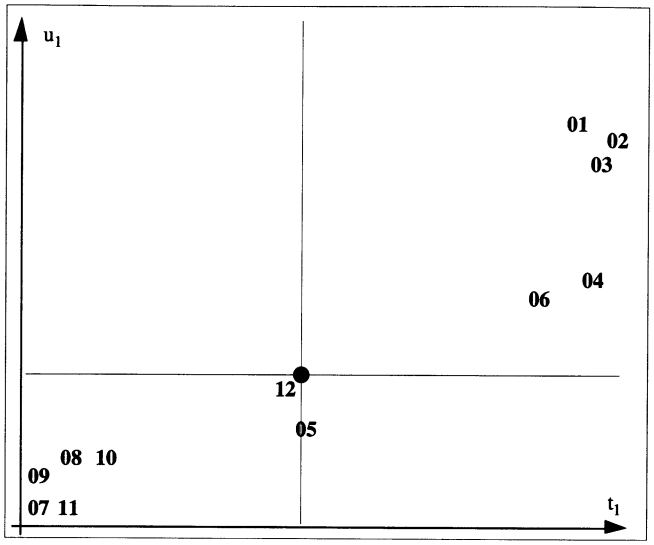


FIGURE 3
Graphique (t_1, u_1)

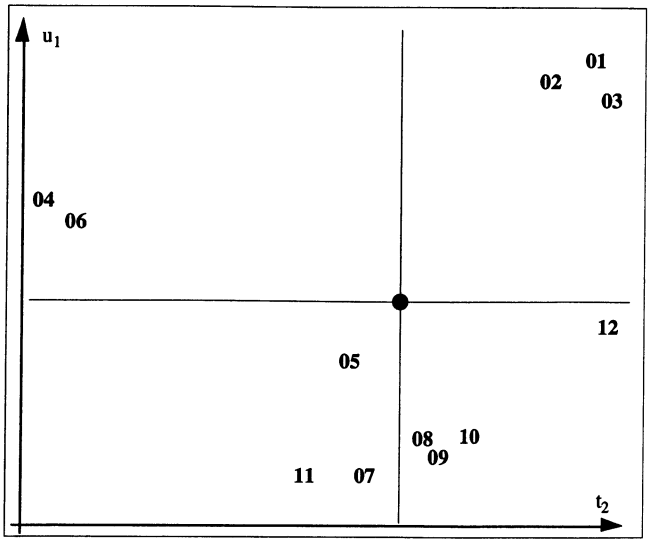


FIGURE 4
Graphique (t_2, u_1)

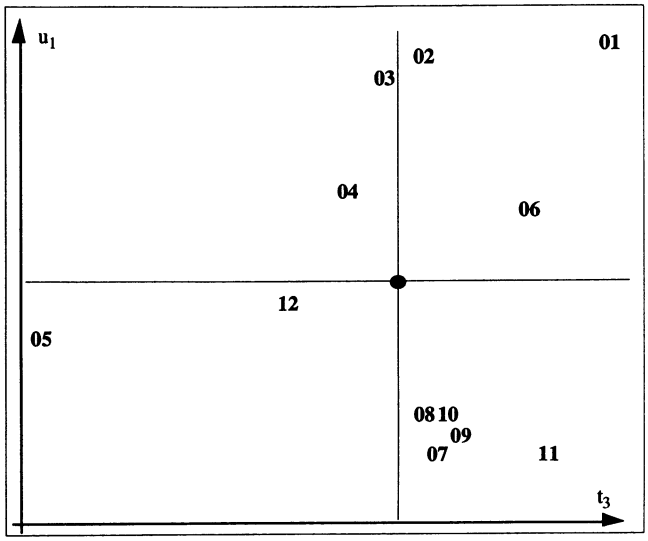


FIGURE 5
Graphique (t_3, u_1)

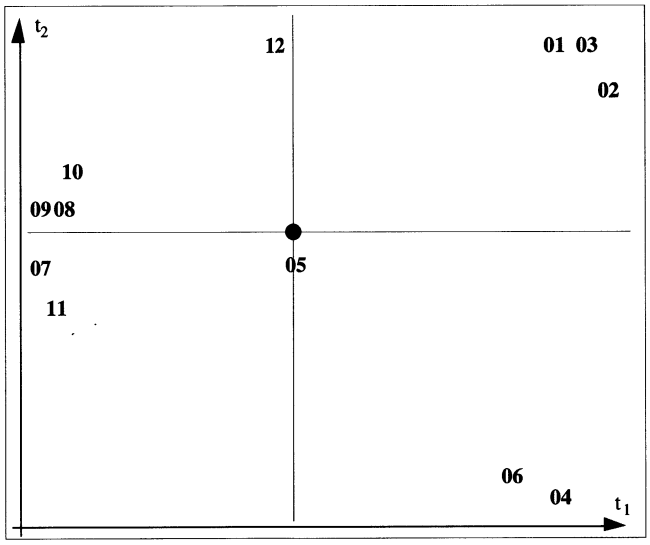


FIGURE 6
Graphique (t_1, t_2)

TABLEAU 4

	Corrélations entre les variables et les composantes PLS	
	t_1	t_2
x_1	−0.906	−0.035
x_2	0.063	−0.845
x_3	−0.906	−0.033
x_4	−0.704	0.234
x_5	0.587	−0.038
x_6	0.920	0.371
x_7	−0.824	0.402
y	0.962	0.230

De $\lambda_1 = \text{Var}\,(t_1) = 3.98$ et $\lambda_2 = \text{Var}\,(t_2) = 0.71$ on déduit que la première colonne du tableau 4 vaut $\sqrt{3.98}\begin{bmatrix} p_1 \\ r_1 \end{bmatrix}$ et la deuxième colonne $\sqrt{0.71}\begin{bmatrix} p_2 \\ r_2 \end{bmatrix}$.

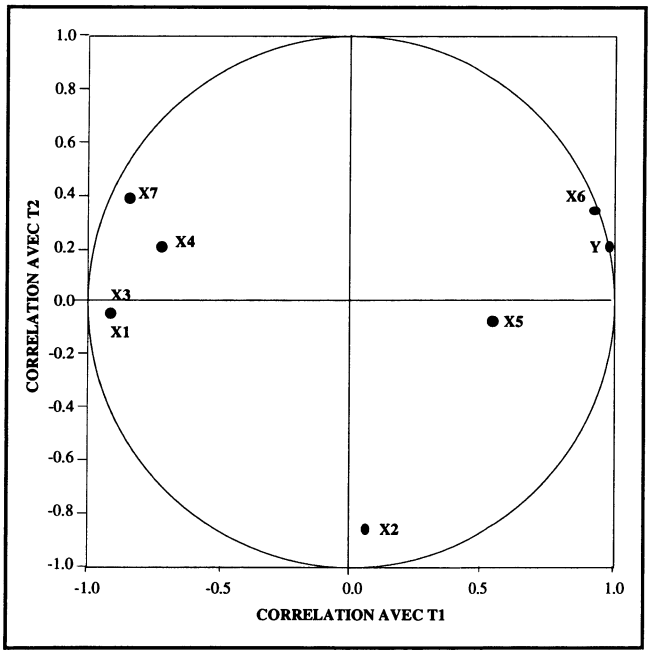


FIGURE 7
Le cercle des corrélations

Ce graphique traduit à la fois les liaisons internes aux variables x_j et entre les x_j et y :

– corrélations positives entre x_1, x_3, x_4, x_7

– corrélation positive entre x_5 et x_6

– y est corrélée positivement à x_5 et x_6 et négativement aux variables x_1, x_3, x_4, x_7.

1.7. Exemple 2

1.7.1. Les données

On trouve dans Jackson [1991] un exemple de régression PLS sur les données suivantes :

Les données ont été collectées par le Dr. A.C. Linnerud de l'Université de Caroline du Nord. Il a mesuré sur vingt hommes d'âge moyen s'entraînant dans un club de gymnastique trois caractéristiques physiques et leurs résultats à trois types d'exercice. Les variables physiques sont les variables explicatives, et les résultats aux exercices les variables à expliquer.

Les données figurent dans le tableau 5.

TABLEAU 5
Données de Linnerud

n° de l'individu	Poids	Tour de taille	Pouls	Tractions à la barre fixe	Flexions	Sauts
1	191	36	50	5	162	60
2	189	37	52	2	110	60
3	193	38	58	12	101	101
4	162	35	62	12	105	37
5	189	35	46	13	155	58
6	182	36	56	4	101	42
7	211	38	56	8	101	38
8	167	34	60	6	125	40
9	176	31	74	15	200	40
10	154	33	56	17	251	250
11	169	34	50	17	120	38
12	166	33	52	13	210	115
13	154	34	64	14	215	105
14	247	46	50	1	50	50
15	193	36	46	6	70	31
16	202	37	62	12	210	120
17	176	37	54	4	60	25
18	157	32	52	11	230	80
19	156	33	54	15	225	73
20	138	33	68	2	110	43
Moy.	178.60	35.40	56.10	9.45	145.55	70.30
Ec-type	24.69	3.20	7.20	5.29	62.57	51.28

Il s'agit de relier les variables y_1 = Tractions, y_2 = Flexions, y_3 = Sauts aux variables explicatives x_1 = Poids, x_2 = Tour de taille, x_3 = Pouls. L'analyse de la matrice des corrélations (tableau 6) montre les points suivants :

– corrélation positive entre Poids et Tour de taille,

– corrélations négatives entre (Poids, Tour de taille) et Pouls,

– corrélations positives entre Tractions, Flexions et Sauts,

– les variables Tractions, Flexions et Sauts sont corrélées négativement à Poids et Tour de taille, et positivement à Pouls.

TABLEAU 6
Données de Linnerud, matrice de corrélations

	Poids x_1	Tour de taille x_2	Pouls x_3	Tractions y_1	Flexions y_2	Sauts y_3
Poids	1	0.8702	−0.3658	−0.3897	−0.4931	−0.2263
Tour de taille		1	−0.3529	−0.5522	−0.6456	−0.1915
Pouls			1	0.1506	0.2250	0.0349
Tractions				1	0.6957	0.4958
Flexions					1	0.6692
Sauts						1

Une régression «cohérente» devrait donc être de la forme $y_k = \beta_0 + \beta_1$ Poids + β_2 Tour de taille + β_3 Pouls + ε avec $\beta_1, \beta_2 < 0$ et $\beta_3 > 0$.

Les estimations $\widehat{\beta}_j$ du tableau 7 des résultats des régressions des variables centrées-réduites Tractions*, Flexions* et Sauts* sur les variables centrées réduites Poids*, Tour de taille*, Pouls* montrent des résultats incohérents.

TABLEAU 7
Données de Linnerud, résultats des régressions

	Tractions*	Flexions*	Sauts*
Poids*	0.3683	0.2872	−0.2590
Tour de taille*	−0.8818	−0.8898	0.0146
Pouls*	−0.0258	0.0161	−0.0546
Ecart-type du résidu	0.8856	0.8180	1.0599
Corrélation multiple R	0.5827	0.6607	0.2322

La régression PLS va permettre d'améliorer ces résultats.

1.7.2. Résultats de la régression PLS

1.7.2.1. Calcul des composantes t_h, u_h

Le logiciel SIMCA permet d'obtenir les vecteurs propres w_h, c_h et les composantes PLS t_h, u_h. Ces résultats sont donnés dans le tableau 8.

TABLEAU 8
Données de Linnerud, résultats de la régression PLS

	w_1	w_2	w_3		c_1	c_2	c_3
Poids	−0.590	0.469	0.657	Tractions	0.614	0.749	0.689
Tour de taille	−0.771	−0.568	−0.287	Flexions	0.747	0.647	0.657
Pouls	0.239	−0.676	0.697	Sauts	0.257	0.145	−0.307

Obs	t_1	u_1	t_2	u_2	t_3	u_3
1	−.643	−.667	.591	−.308	.131	−.411
2	−.770	−2.406	.167	−2.296	−.134	−1.431
3	−.907	−.148	−.521	1.073	−.048	.451
4	.688	−.637	−.680	−1.175	−.346	.086
5	−.487	.831	1.133	1.850	.182	.585
6	−.229	−2.344	−.072	−2.643	−.025	−1.260
7	−1.404	−1.547	−.077	.019	.572	.317
8	.744	−1.431	−.211	−2.654	.032	−1.074
9	1.715	2.051	−.655	.728	1.557	1.36
10	1.163	5.448	.167	4.526	−.333	.590
11	.365	.735	.701	1.145	−.201	.659
12	.743	2.522	.698	1.981	−.002	.354
13	1.187	2.748	−.757	1.809	−.336	1.252
14	−4.390	−3.990	−.760	.341	−.255	.525
15	−.823	−2.691	.974	−2.074	.083	−1.468
16	−.749	2.359	−.521	3.509	.667	1.832
17	−.393	−3.376	−.203	−3.494	−.564	−1.548
18	1.199	2.220	.783	1.037	−.092	.228
19	1.049	2.883	.373	2.321	−.319	1.304
20	1.942	−2.559	−1.129	−5.696	−.568	−2.350

Dans SIMCA chaque composante PLS u_h est normalisée de manière à ce que le coefficient de régression de u_h sur t_h soit égal à 1. On trouve donc dans le tableau 8 les composantes PLS u_h définies par

$$u_h = \frac{1}{b_h} \cdot F_{h-1} c_h$$

où $b_h = \dfrac{t_h' F_{h-1} c_h}{t_h' t_h'}$ est le coefficient de régression de $F_{h-1}c_h$ sur t_h.

Les résultats des régressions

$$E_0 = t_1 p_1' + t_2 p_2' + t_3 p_3'$$
$$F_0 = t_1 r_1' + t_2 r_2' + t_3 r_3' + F_3$$

sont donnés dans le tableau 9.

TABLEAU 9

Données de Linnerud, coefficients des régressions de E_0 et F_0 sur t_1, t_2, t_3

	p_1	p_2	p_3		r_1	r_2	r_3
Poids	−0.666	−0.020	0.657	Tractions	0.342	0.336	0.477
Tour de taille	−0.676	−0.355	−0.287	Flexions	0.416	0.291	0.455
Pouls	0.359	−1.194	0.697	Sauts	0.143	0.065	−0.213

De $r_h = \dfrac{F_{h-1}' t_h}{t_h' t_h} = \dfrac{F_{h-1}' E_{h-1} w_h}{t_h' t_h} = \dfrac{t_h' F_{h-1} c_h}{t_h' t_h} c_h = b_h c_h$

on déduit

$$||r_h||^2 = b_h^2$$

et

$$u_h = \frac{1}{||r_h||^2} F_{h-1} r_h$$

Les composantes PLS u_h pouvant s'exprimer directement en fonction des r_h, le logiciel SIMCA ne fournit pas les vecteurs propres c_h.

1.7.2.2. Valeurs explicatives des composantes PLS

Les formules (29) et (30) s'écrivent :

$$||E_0||^2 = ||t_1||^2 \cdot ||p_1||^2 + ||t_2||^2 \cdot ||p_2||^2 + ||t_3||^2 \cdot ||p_3||^2$$

et

$$||F_0||^2 = ||t_1||^2 \cdot ||r_1||^2 + ||t_2||^2 \cdot ||r_2||^2 + ||t_3||^2 \cdot ||r_3||^2 + ||F_3||^2.$$

On a :

$$\begin{aligned}
||E_0||^2 &= ||F_0||^2 = 3(N-1) = 57 \\
||t_1||^2 &= 38.47 \\
||t_2||^2 &= 8.33 \\
||t_3||^2 &= 4.47 \\
||p_1||^2 &= 1.029 \qquad ||r_1||^2 = 0.310 \\
||p_2||^2 &= 1.552 \qquad ||r_2||^2 = 0.20 \\
||p_3||^2 &= 1.000 \qquad ||r_3||^2 = 0.48
\end{aligned}$$

d'où le tableau 10 donnant les variances expliquées par les composantes t_h.

TABLEAU 10
Pouvoir explicatif des t_h

	Composantes		
	t_1	t_2	t_3
% de variance de X expliqué	69,45	22,70	7,85
% de variance de Y expliqué	20,92	2,92	3,76

Ainsi les données étudiées sont essentiellement unidimensionnelles, au niveau de la relation entre Y et X.

1.7.2.3. Interprétation des composantes PLS

Les composantes PLS t_h peuvent s'écrire en fonction des X_j centrées-réduites :

$$\begin{aligned}
t_1 &= E_0 w_1 = -0.590 \text{ Poids}^* - 0.771 \text{ Tour de taille}^* + 0.239 \text{ Pouls} \\
t_2 &= E_1 w_2 = E_0(I - w_1 p_1')w_2 = E_0 \widetilde{w}_2 \\
&= 0.368 \text{ Poids}^* - 0.700 \text{ Tour de taille}^* - 0.636 \text{ Pouls} \\
t_3 &= E_2 w_3 = E_0(I - w_1 p_1')(I - w_2 p_2')w_3 = E_0 \widetilde{w}_3 \\
&= 0.935 \text{ Poids}^* - 0.802 \text{ Tour de taille}^* + 0.223 \text{ Pouls}^*
\end{aligned}$$

On donne dans le tableau 11 les corrélations de chaque X_j avec les t_h.

TABLEAU 11
Corrélations entre les variables et les composantes PLS

	t_1	t_2	t_3
Poids	−0.95	−0.01	0.32
Tour de taille	−0.96	−0.23	−0.14
Pouls	0.51	−0.79	0.34
Traction	0.49	0.22	0.23
Flexion	0.59	0.19	0.22
Saut	0.20	0.04	−0.10

u_1	u_2	u_3
0.88	0.78	0.76
0.94	0.71	0.61
0.74	0.69	0.41

On déduit donc que la composante t_1 oppose la variable Pouls aux variables Poids et Tour de taille. Elle classe les individus des grands/gros aux petits/maigres. La deuxième composante t_2 est essentiellement corrélée à la variable Pouls.

La dernière composante t_3 n'apporte plus d'information.

La composante PLS $u_1 = F_0 r_1 / ||r_1||^2$ s'écrit :

$$u_1 = 1.103 \text{ Tractions}^* + 1.342 \text{ Flexions}^* + 0.461 \text{ Sauts}^*$$

La composante u_1 est un indice de performance donnant plus de poids aux variables Tractions et Flexions qu'à la variable Sauts. On a cor $(u_1, t_1) = 0.55$.

La composante u_2 s'écrit :

$$u_2 = 1.68F_{11} + 1.455F_{12} + 0.325F_{13}$$

où F_{1k} est le vecteur des résidus de la régression de F_{0k} sur t_1. Cette composante u_2 ne peut s'exprimer en fonction des variables y_k, mais on peut utiliser les corrélations entre les Y_k et u_2 du tableau 11 pour l'interpréter. La composante u_2 est redondante avec u_1. On a cor $(u_1, u_2) = 0.83$. Il en va de même pour la composante u_3 : cor $(u_1, u_3) = 0.72$ et cor $(u_2, u_3) = 0.90$.

1.7.2.4. Les graphiques

Le plan (t_1, t_2) de la figure 8 est analogue à un premier principal dans l'espace des X_j, mais orienté vers l'explication des Y_k. Cette carte des individus montre clairement la position particulière de l'individu 14. Il est très gros (Poids = 247, Tour de taille = 46) et obtient les plus faibles résultats du groupe en Traction (= 1) et Flexion (= 50).

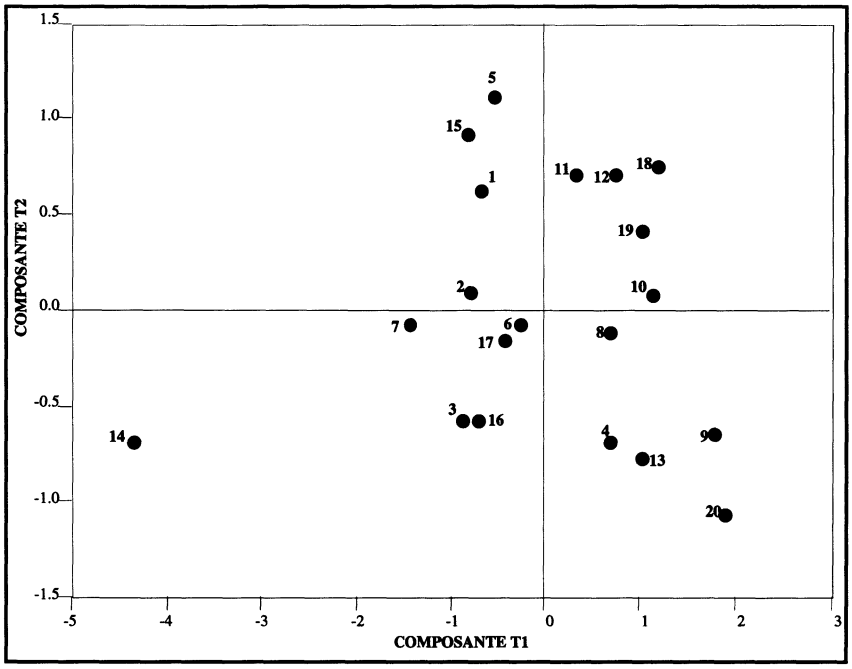


FIGURE 8
Plan des composantes (t_1, t_2)

Le cercle des corrélations de la figure 9 traduit bien les corrélations internes à chaque groupe de variables X, Y comme les corrélations entre les groupes X et Y.

Le graphique de la figure 10 met en évidence les rôles particuliers joués par les individus 10, 14 et 20 :

– les individus 10 et 20 ont à peu près les mêmes caractéristiques physiques et des résultats aux exercices opposés : l'individu 10 obtient les meilleurs scores aux exercices Tractions, Flexions et Sauts sur l'ensemble des 20 individus, alors que l'individu 20 est parmi les plus faibles,

– l'individu 14 a déjà été mis en évidence dans la figure 8.

Il faudrait refaire l'analyse en mettant en points supplémentaires ces trois individus.

On peut également montrer les graphiques des régressions de Flexions, Tractions et Sauts sur t_1.

Ces 3 graphiques confirment les rôles particuliers des individus 14 et 20 pour la prédiction des variables Tractions et Flexions. L'individu 10 a un score en Sauts (=250) exceptionnel.

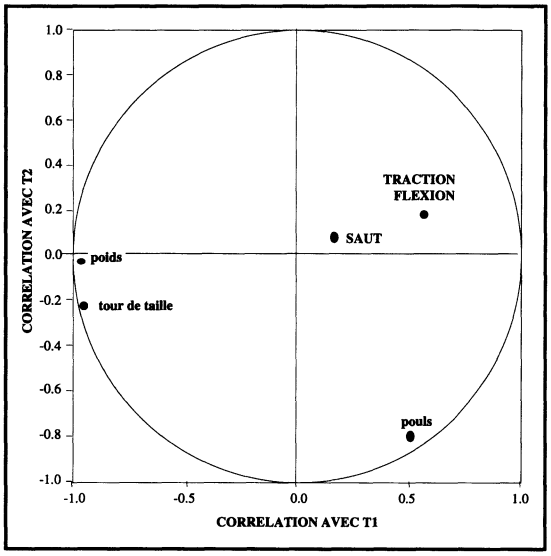


FIGURE 9
Le cercle des corrélations

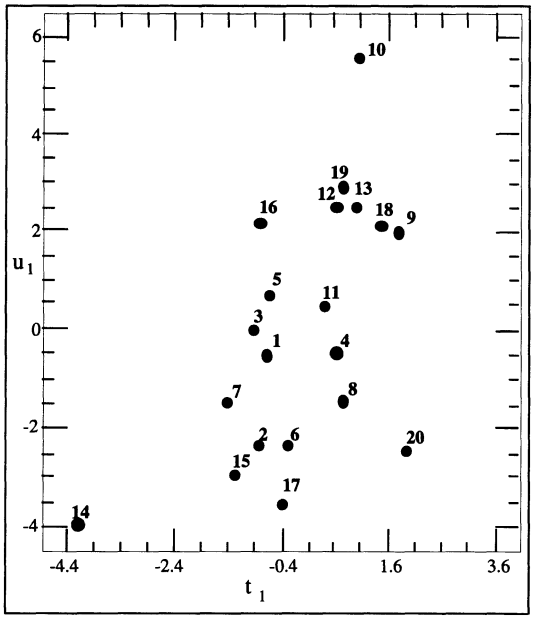


FIGURE 10
Graphique (u_1, t_1)

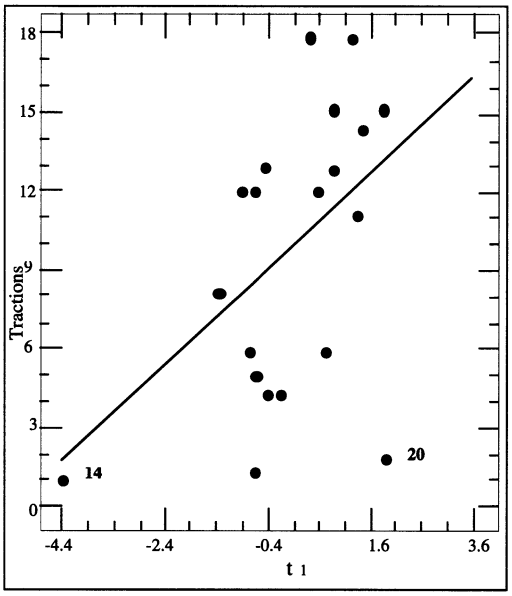


FIGURE 11
Régression de Tractions sur t_1

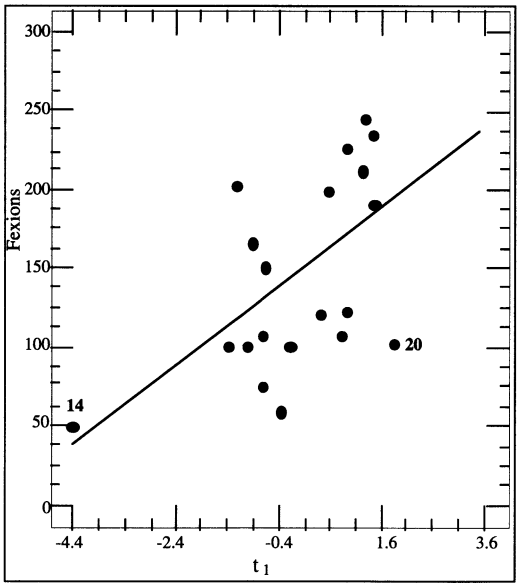


FIGURE 12
Régression de Flexions sur t_1

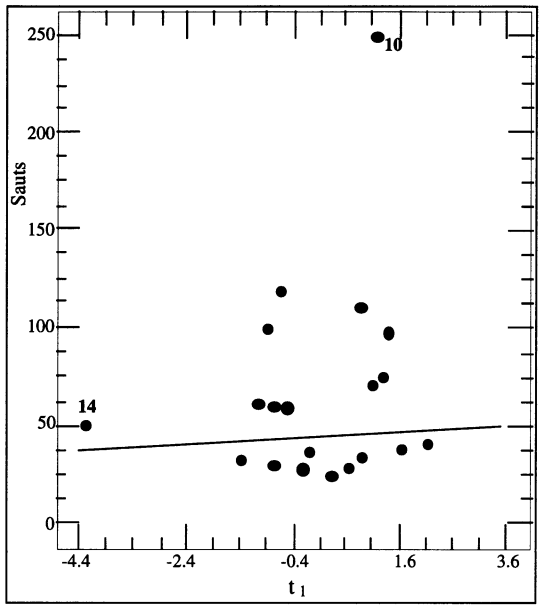


FIGURE 13
Régression de Sauts sur t_1

1.7.2.5. Les équations de régression PLS

Seule la composante t_1 explique une part de variance de Y_k significative. Nous avons donc réalisé les régressions simples de Tractions, Flexions et Sauts sur la composante t_1.

$$\begin{array}{lcll} \text{Tractions} & \approx & 9.45 + 1.81t_1, & \text{cor (Tractions}, t_1) = .49 \\ \text{Flexions} & \approx & 145.56 + 26.03t_1, & \text{cor (Flexions}, t_1) = .59 \\ \text{Sauts} & \approx & 70.30 + 7.31t_1, & \text{cor (Sauts}, t_1) = .20 \end{array}$$

La composante t_1 s'écrivant :

$$t_1 = -.590\left[\frac{\text{Poids} - 178.6}{24.09}\right] - .771\left[\frac{\text{Tour de taille} - 35.4}{3.2}\right] + .239\left[\frac{\text{Pouls} - 56.1}{7.21}\right]$$

on peut revenir aux données d'origine :

$$\begin{array}{lcl} \text{Tractions} & \approx & 29.12 - 0.0429\ \text{Poids} - 0.434\ \text{Tour de taille} + 0.0597\ \text{Pouls} \\ \text{Flexions} & \approx & 429.78 - 0.62\ \text{Poids} - 6.26\ \text{Tour de taille} + 0.86\ \text{Pouls} \\ \text{Sauts} & \approx & 150.1 - 0.17\ \text{Poids} - 1.76\ \text{Tour de taille} + 0.24\ \text{Pouls} \end{array}$$

et en données centrées réduites :

$$\begin{array}{lcl} \text{Tractions}^* & \approx & -0.20\ \text{Poids}^* - 0.26\ \text{Tour de taille}^* + 0.08\ \text{Pouls}^* \\ \text{Flexions}^* & \approx & -0.24\ \text{Poids}^* - 0.32\ \text{Tour de taille} + 0.10\ \text{Pouls}^* \\ \text{Sauts}^* & \approx & -0.08\ \text{Poids}^* - 0.11\ \text{Tour de taille}^* + 0.03\ \text{Pouls}^* \end{array}$$

La comparaison avec les résultats des régressions multiples ordinaires (Tableau 7) est instructive : on a rendu les équations de régressions cohérentes en acceptant une légère diminution des pouvoirs explicatifs de chaque régression.

1.8. Conclusion

La régression PLS permet de rapprocher le statisticien du chercheur dans les problèmes de modélisation. En général ce dernier souhaite en effet conserver dans son modèle toutes les variables importantes tout en obtenant des équations de régression cohérentes. Lorsque, en régression multiple il y a multicolinéarité et/ou un nombre important de variables explicatives par rapport au nombre d'observations, la solution la plus courante consiste à exclure des variables explicatives par des méthodes de pas-à-pas. La régression PLS permet dans ces situations de conserver toutes les variables explicatives tout en obtenant une équation de régression cohérente.

Lorsqu'il s'agit de relier un ensemble de Y_k à un ensemble de X_j, on peut essayer l'analyse canonique. L'intérêt pratique de cette méthode est souvent faible car elle ne fait le plus souvent qu'isoler des couples (X_j, Y_k) à forte corrélation. Ainsi les variables canoniques expliquent une part qui peut être faible des variances des groupes de variables X et Y.

La régression PLS apparaît comme un compromis entre l'analyse canonique des groupes de variables X et Y et des analyses en composantes principales de chacun de ces groupes de variables.

2. Applications

Nous allons présenter dans cette deuxième partie deux applications réalisées chez Rhône-Poulenc, la première au Centre de Recherches d'Aubervilliers et la seconde au Centre d'Industrialisation de Décines.

Les données relatives à ces applications ont la particularité de contenir des variables qualitatives. La prise en compte de variables qualitatives en régression PLS ne pose pas de problèmes particuliers. Après avoir étudié en détail les deux applications nous proposerons en conclusion une méthodologie générale de la régression PLS qualitative.

2.1. Un problème de formulation d'une huile silicone fonctionnalisée

2.1.1. Description du problème

Il s'agit d'optimiser les niveaux de cinq facteurs expérimentaux gouvernant la stabilité d'une micro-émulsion d'une huile silicone fonctionnalisée. Les facteurs X_1 à X_4 représentent des quantités de constituants actifs et prennent les niveaux $-1, 0, 1$ et le facteur X_5 est une condition de fabrication codée -1 ou 1. On a également introduit toutes les interactions $X_h \times X_k$ et les carrés $X_1^2, \ldots, X_4^2$. Ainsi l'ensemble X des variables explicatives est constitué de 19 variables :

$$X = \{X_1, \ldots, X_5, X_1^2, \ldots, X_4^2, X_1 \times X_2, \ldots, X_4 \times X_5\}.$$

Un plan d'expérience D-optimal (Fedorov, 1972) a permis d'obtenir 58 expériences décrites dans le tableau 12.

Pour chaque expérience l'aspect de la formulation a été noté en sept instants différents. Le résultat Y_k de l'observation à l'instant k se traduit par l'attribution d'une note (1, 2 ou 3) correspondant à une description totalement qualitative de l'aspect de la formulation :

- 1 = système polyphasique (SP), la formulation s'est dégradée,
- 2 = système fluide transparent (SFT), la formulation est stable, ne se dégrade pas,
- 3 = système dit "tout système visqueux" (TSV), la formulation n'est pas complètement dégradée mais n'est pas conforme.

Après entretien avec le chercheur, il est clair que cette variable réponse n'est pas ordinale. Cette variable sera donc considérée par la suite comme nominale à 3 modalités. Le tableau des résultats figure ci-après (tableau 13).

Il s'agit donc de relier les réponses qualitatives aux sept instants $Y = \{Y_1, \ldots, Y_7\}$ aux conditions de fabrication décrites par $X = \{X_1, \ldots, X_5, X_1^2, \ldots, X_4^2, X_1 \times X_2, \ldots, X_4 \times X_5\}$.

2.1.2 La méthodologie statistique

On cherche à établir une équation de régression expliquant la stabilité de l'émulsion en fonction des facteurs expérimentaux. Pour ce faire nous avons adopté la méthodologie suivante :

a) On réalise une analyse des correspondances multiples du tableau Y des réponses aux 7 instants. On ne retient que les composantes principales utiles à l'interprétation des résultats.

b) On réalise des analyses de la variance univariées de chaque composante principale retenue en a) sur les conditions de fabrication décrites par X. On ne retient que les variables les plus explicatives.

c) On calcule la régression PLS des composantes principales retenues en a) sur les facteurs explicatifs retenus en b).

TABLEAU 12
Le plan d'expériences

N°exp.	X_1	X_2	X_3	X_4	X_5
1	−1	−1	−1	−1	−1
2	−1	−1	−1	1	1
3	−1	−1	1	−1	1
4	−1	−1	1	1	−1
5	−1	0	−1	1	−1
6	−1	0	0	−1	1
7	−1	1	−1	−1	1
8	−1	1	−1	0	−1
9	−1	1	0	1	−1
10	−1	1	1	−1	−1
11	−1	1	1	1	1
12	0	−1	0	1	1
13	0	0	1	−1	−1
14	0	1	−1	1	−1
15	0	1	1	0	1
16	1	−1	−1	−1	1
17	1	−1	−1	1	−1
18	1	−1	0	0	−1
19	1	−1	1	−1	−1
20	1	−1	1	1	1
21	1	0	−1	0	1
22	1	1	−1	−1	−1
23	1	1	−1	1	1
24	1	1	1	−1	1
25	1	1	1	1	−1
26	0	0	0	0	−1
27	0	0	0	0	−1
28	0	0	0	0	1
29	0	0	0	0	1
30	−1	−1	−1	−1	−1
31	−1	−1	−1	1	1
32	−1	−1	1	−1	1
33	−1	−1	1	1	−1
34	−1	0	−1	1	−1
35	−1	0	0	−1	1
36	−1	1	−1	−1	1
37	−1	1	−1	0	−1
38	−1	1	0	1	−1
39	−1	1	1	−1	−1
40	−1	1	1	1	1
41	0	−1	0	1	1
42	0	0	1	−1	−1
43	0	1	−1	1	−1
44	0	1	1	0	1
45	1	−1	−1	−1	1
46	1	−1	−1	1	−1
47	1	−1	0	0	−1
48	1	−1	1	−1	−1
49	1	−1	1	1	1
50	1	0	−1	0	1
51	1	1	−1	−1	−1
52	1	1	−1	−1	1
53	1	1	1	−1	1
54	1	1	1	1	−1
55	0	0	0	0	−1
56	0	0	0	0	−1
57	0	0	0	0	1
58	0	0	0	0	1

TABLEAU 13
Observations à 7 époques

N°	t1	t2	t3	t4	t5	t6	t7	n°	t1	t2	t3	t4	t5	t6	t7
1	2	3	2	2	1	1	1	30	3	3	3	2	1	1	1
2	3	3	3	3	2	2	2	31	3	3	3	3	2	2	2
3	1	1	1	1	1	1	1	32	1	3	1	1	1	1	1
4	1	1	1	1	1	1	1	33	1	1	1	1	1	1	1
5	3	3	3	3	1	1	1	34	3	3	3	1	1	1	1
6	1	1	1	1	1	1	1	35	1	1	1	1	1	1	1
7	3	3	3	3	1	1	1	36	3	3	3	3	1	1	1
8	3	3	3	3	3	1	1	37	3	3	3	3	1	1	1
9	1	1	1	1	1	1	1	38	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	39	1	1	1	1	1	1	1
11	1	1	1	1	1	1	1	40	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	41	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	42	2	2	2	2	2	2	2
14	3	3	1	3	3	3	1	43	3	3	1	3	3	1	1
15	1	1	1	1	1	3	2	44	1	1	1	1	2	1	2
16	1	1	1	1	2	2	2	45	3	3	2	2	2	2	2
17	1	1	1	1	1	1	1	46	3	3	1	1	1	1	1
18	1	1	1	1	1	1	1	47	1	1	1	1	1	1	1
19	1	1	3	1	1	1	1	48	1	1	1	1	1	1	1
20	1	1	1	1	1	1	1	49	1	1	1	1	1	1	1
21	3	3	3	3	1	1	1	50	3	3	3	3	1	1	1
22	3	3	3	3	3	3	3	51	3	3	3	3	3	1	3
23	3	3	3	3	3	3	1	52	3	3	3	3	2	1	3
24	1	1	1	1	1	1	1	53	1	1	1	1	1	1	1
25	1	1	1	1	2	1	1	54	1	2	2	2	2	3	1
26	1	1	1	1	1	1	1	55	1	1	1	1	1	1	1
27	1	1	1	1	1	1	1	56	1	1	1	1	1	1	1
28	1	1	1	1	1	1	1	57	1	1	1	1	1	1	1
29	1	1	1	1	1	1	1	58	1	1	1	1	1	1	1

d) On détermine enfin les niveaux des facteurs explicatifs favorisant les produits qualifiés SFT.

2.1.2.1. L'analyse des correspondances multiples

Les principaux résultats de l'ACM sont :

– Le tableau des valeurs propres (tableau 14) montre qu'il y a quatre valeurs propres supérieures à $1/7 = 0.14$. La structure recherchée étant mise en évidence dans le premier plan principal nous n'avons conservé dans cette étude que les deux premières composantes principales.

TABLEAU 14
Décomposition de l'inertie

numéro	valeur propre	%	% cumulé
1	0,6466	32,33	32,33
2	0,5158	25,79	58,12
3	0,2386	11,93	70,05
4	0,1873	9,36	79,41
5	0,1129	5,65	85,06
6	0,0874	4,37	89,43
7	0,0587	2,93	92,36
8	0,0522	2,61	94,97
9	0,0293	1,46	96,44
10	0,0229	1,14	97,58
11	0,0183	0,91	98,49
12	0,0142	0,71	99,21
13	0,0104	0,52	99,73
14	0,0055	0,27	100,00

– Nous donnons dans le tableau 15 les coordonnées des points-modalités et la carte des modalités dans la figure 14. Cette carte est particulièrement informative puisque les formulations y sont nettement séparées selon leur aspect.

TABLEAU 15
Coordonnées des points-modalités

n°point	modalité	axe 1	axe 2
1	SP	0,64	−0,27
2	SFT	−1,91	−3,01
3	TSV	−1,05	0,83
4	SP	0,71	−0,21
5	SFT	−2,10	−3,31
6	TSV	−0,98	0,67
7	SP	0,61	−0,12
8	SFT	−1,90	−2,62
9	TSV	−0,97	0,95
10	SP	0,66	−0,16
11	SFT	−1,68	−2,06
12	TSV	−1,11	1,10
13	SP	0,44	0,02
14	SFT	−1,27	−1,14
15	TSV	−1,23	1,55
16	SP	0,30	0,12
17	SFT	−1,77	−1,53
18	TSV	−1,11	0,36
19	SP	0,28	0,08
20	SFT	−1,23	−1,28
21	TSV	−1,57	1,69

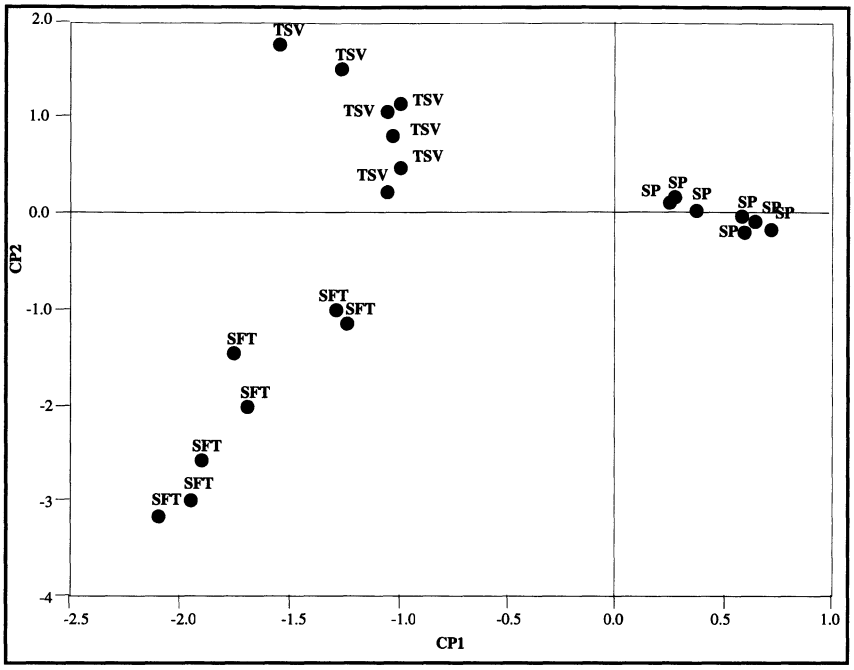


FIGURE 14
Carte des modalités

A propos du tableau 15 nous précisons que les trois premiers points correspondent aux modalités SP, SFT, TSV à l'époque 1, les trois suivants correspondant à ces mêmes modalités à l'époque 2 ... et les trois derniers à ces modalités à l'époque 7.

– Les résultats concernant les points-individus (les expériences) sont donnés dans le tableau 16 et visualisés dans la figure 15.

Remarque : le point 18 cache les points n°3, 4, 6, 9, 10, 11, 12, 13, 17, 18, 20, 24, 26, 27, 28, 29, 33, 35, 38, 39, 40, 41, 47, 48, 49, 53, 55, 56, 57, 58; tous ces points correspondent à une expérience à réponses SP aux sept époques.

La carte des expériences de la figure 15 montre la difficulté du problème à résoudre. Le bon quadrant (produits SFT) comporte peu d'expériences, et la seule expérience entièrement satisfaisante (expérience n°42) apparaît comme isolée. Elle est associée à des valeurs des deux premières composantes principales minimales.

La première composante principale CP_1 sépare la formulation dégradée (SP) des autres cas. La deuxième composante principale CP_2 sépare ensuite la situation intermédiaire (TSV) de la formulation recherchée (SFT).

On recherche donc des conditions de fabrication rendant aussi négatives que possible les deux premières composantes principales.

TABLEAU 16
Coordonnées des points-individus (expériences)

n°point	axe 1	axe 2	n°point	axe 1	axe 2
1	−0.97	−1.35	30	−0.65	0.12
2	−1.49	−0.08	31	−1.49	−0.08
3	0.65	−0.11	32	0.34	0.07
4	0.65	−0.11	33	0.65	−0.11
5	−0.55	0.75	34	−0.24	0.50
6	0.65	−0.11	35	0.65	−0.11
7	−0.55	0.75	36	−0.55	0.75
8	−0.85	1.06	37	−0.55	0.75
9	0.65	−0.11	38	0.65	−0.11
10	0.65	−0.11	39	0.65	−0.11
11	0.65	−0.11	40	0.65	−0.11
12	0.65	−0.11	41	0.65	−0.11
13	0.65	−0.11	42	−2.11	−2.97
14	−0.82	0.89	43	−0.56	0.84
15	0.13	−0.33	44	0.07	−0.61
16	−0.29	−0.94	45	−1.75	−1.42
17	0.65	−0.11	46	0.05	0.29
18	0.65	−0.11	47	0.65	−0.11
19	0.36	0.11	48	0.65	−0.11
20	0.65	−0.11	49	0.65	−0.11
21	−0.55	0.75	50	−0.55	0.75
22	−1.43	1.42	51	−1.17	1.38
23	−1.10	1.10	52	1.18	0.84
24	0.65	−0.11	53	0.65	−0.11
25	0.34	−0.34	54	−1.27	−1.78
26	0.65	−0.11	55	0.65	−0.11
27	0.65	−0.11	56	0.65	−0.11
28	0.65	−0.11	57	0.65	−0.11
29	0.65	−0.11	58	0.65	−0.11

2.1.2.2. Analyse de la variance des composantes principales sur les conditions de fabrication

Les analyses de la variance univariées des composantes principales CP_1, CP_2 sur les conditions de fabrication ont permis de détecter des termes particulièrement explicatifs :

– X_3, X_3^2, $X_1 \times X_2$, $X_1 \times X_3$, $X_2 \times X_3$, $X_3 \times X_4$, $X_3 \times X_5$ pour CP_1

– X_3, $X_2 \times X_3$ pour CP_2.

Les résultats de ces analyses de la variance apparaissent dans le tableau 17.

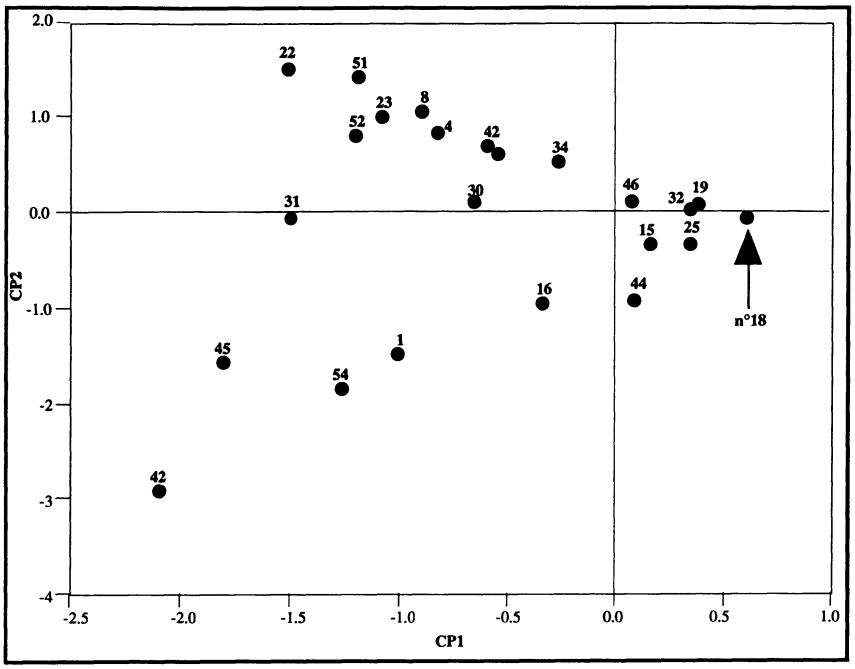


FIGURE 15
Carte des individus

TABLEAU 17
Analyses de la variance sur les conditions de fabrication (variables significatives)

Variable dépendante	Facteur	F	Niveau de signification
CP_1	X_3	35.63	0.0001
	X_3^2	18.29	0.0001
	$X_1 \times X_2$	3.33	0.04
	$X_1 \times X_3$	3.80	0.03
	$X_2 \times X_3$	2.95	0.06
	$X_3 \times X_4$	10.44	0.0001
	$X_3 \times X_5$	4.32	0.02
CP_2	X_3	7.63	0.001
	$X_2 \times X_3$	9.98	0.002

2.1.2.3. Régression PLS des composantes principales CP_1 et CP_2 sur les variables explicatives significatives

Il est essentiel pour le chercheur de conserver dans un modèle prédictif toutes les variables explicatives significatives. La régression multiple pas à pas descendante

conduit aux modèles

$$CP_1 \approx 0.61 + 0.53X_3 - 0.84X_3^2 - 0.22X_1 \times X_2 + 0.15X_2 \times X_4 + 0.18X_3 \times X_4 \quad (31)$$

$$CP_2 \approx -0.03 + 0.24X_2 - 0.33X_3 - 0.41X_2 \times X_3 - 0.19X_2 \times X_5 + 0.16X_3 \times X_5 \quad (32)$$

et la régression PLS aux équations de régression

$$CP_1 \approx 0.6102 + 0.4974X_3 - 0.8549X_3^2 - 0.2817X_1 \times X_2 - 0.0815X_1 \times X_3 - 0.0033X_2 \times X_3 - 0.0883X_3 \times X_4 + 0.1630X_3 \times X_5 \quad (33)$$

$$CP_2 \approx -0.0965 - 0.3397X_3 - 0.4309X_2 \times X_3 \quad (34)$$

Les signes des coefficients de la régression PLS sont plus cohérents que ceux de la régression pas à pas descendante.

Les équations (33) et (34) sont construites à partir de deux composantes PLS t_1 et t_2 expliquant 30% de la variance du tableau des conditions de fabrication significatives et 57% du tableau des Y_k.

Les reconstitutions des composantes principales CP_1, CP_2 par les équations de régression PLS (33) et (34) apparaissent sur les figures 16 et 17.

Le cercle des corrélations est donné dans la figure 18.

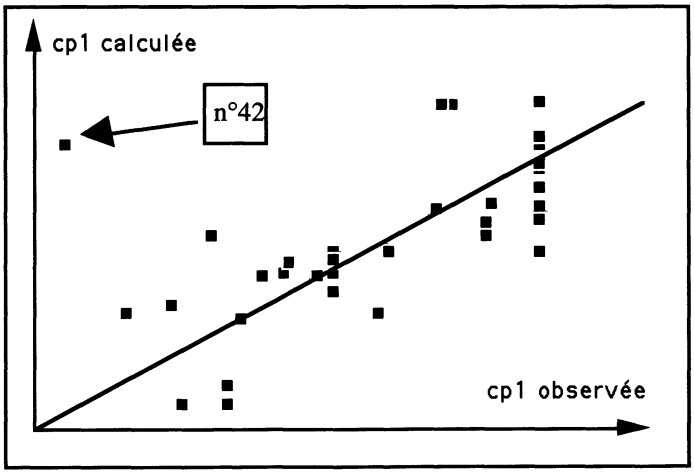


FIGURE 16
Graphique [cp_1 calculée, cp_1 observée]

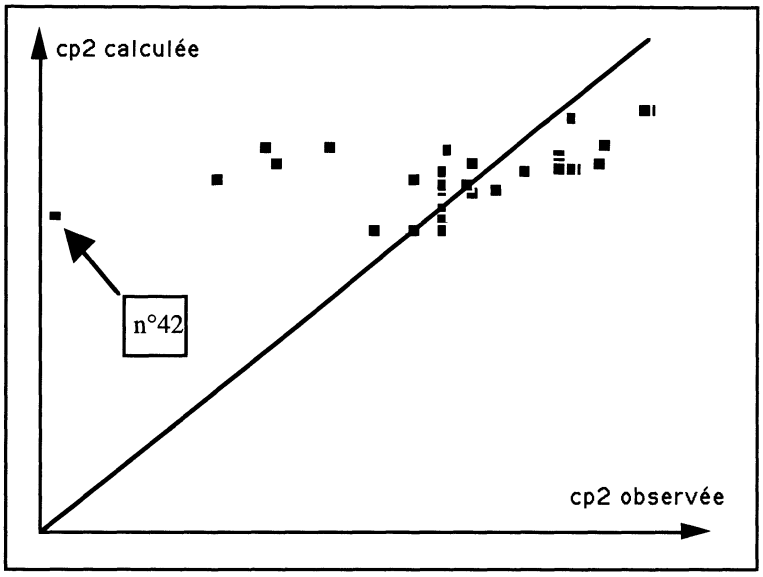


FIGURE 17
Graphique [cp_2 calculée, cp_2 observée]

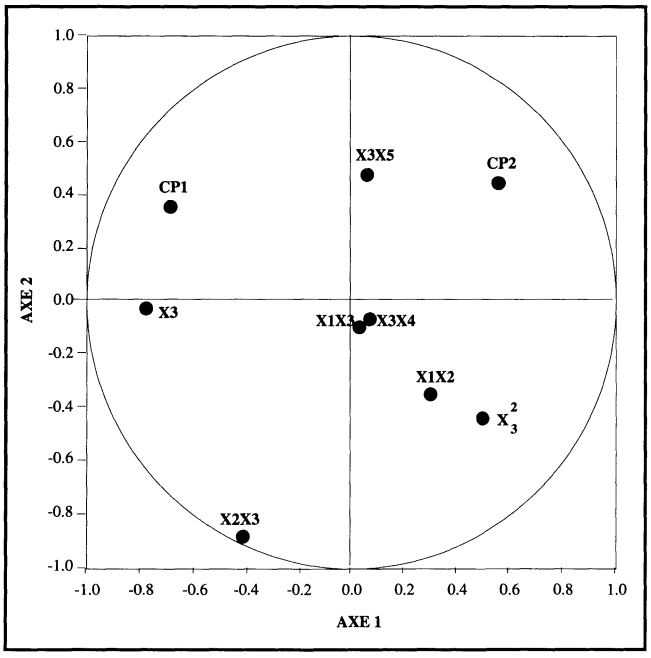


FIGURE 18
Le cercle des corrélations

2.1.2.4. Détermination des conditions optimales

On cherche à minimiser CP1 et CP2 pour être dans le quadrant où se trouvent les 7 modalités SFT. En examinant les équations de régression PLS et le cercle des corrélations, on constate que la configuration permettant de minimiser les composantes principales est déterminée pour X_1, X_2, X_3 et X_4 fixés à -1 et pour X_5 fixé à 1. Dans ces conditions CP_1 vaut -1.35, ce qui est satisfaisant, mais CP_2 n'est que faiblement négative, à savoir $CP_2 = -0.19$. En conclusion le chimiste est capable (ceci a été confirmé expérimentalement) d'éviter la fabrication de produits d'aspect SP, très rhédibitoire. Toutefois, il est encore difficile de savoir comment orienter l'aspect plutôt vers SFT que TSV en jouant sur les facteurs expérimentaux pris en compte dans cette étude. Ainsi la modélisation obtenue ne permet pas de reproduire la qualité de l'expérience n°42.

2.2. Application de la régression PLS à la spectrométrie infrarouge

2.2.1. Introduction

Les spécifications d'un produit fini peuvent correspondre à des concentrations (pureté, impuretés...), à des données de tests physico-chimiques et plus particulièrement à des résultats de tests d'application.

Ces résultats se présentent généralement sous la forme :

– test correct,

– test incorrect.

Quand la nature du test le permet, on peut souhaiter préciser ces données de la façon suivante :

– test correct,

– test incorrect (valeurs inférieures),

– test incorrect (valeurs supérieures).

Ce test qui fournit un résultat très important sur la qualité du produit fini et son orientation future, a parfois l'inconvénient d'être long à mettre en œuvre.

Il est donc intéressant de chercher à développer une méthode de remplacement plus rapide. Dans le cas que nous présentons, la spectrométrie IR et la régression PLS permettent d'obtenir des résultats similaires au test d'application en cours avec un gain de temps très important.

On trouvera une présentation détaillée du contexte de cette application dans Ménardo (1993).

2.2.2. Présentation des données

Nous disposons de vingt échantillons dont les résultats du test d'application sont connus. Pour chaque échantillon, l'information est du type :

– incorrect pour l'application (valeurs inférieures) → *Groupe 1*
– correct pour l'application → *Groupe 2*
– incorrect pour l'application (valeurs supérieures) → *Groupe 3*

La matrice Y est construite de la façon suivante :

N° échantillon	Groupe origine	Variable Y_1	Variable Y_2	Variable Y_3
1	2	0	1	0
2	1	1	0	0
3	1	1	0	0
4	1	1	0	0
5	1	1	0	0
6	1	1	0	0
7	3	0	0	1
8	1	1	0	0
9	3	0	0	1
10	2	0	1	0
11	2	0	1	0
12	3	0	0	1
13	3	0	0	1
14	2	0	1	0
15	2	0	1	0
16	2	0	1	0
17	3	0	0	1
18	3	0	0	1
19	2	0	1	0
20	2	0	1	0

La variable indicatrice Y_j vaut 1 pour les individus du groupe j et 0 pour les autres.

Pour chaque échantillon, le spectre IR a été enregistré sur le même spectromètre IRTF Bruker IFS48, dans les mêmes conditions.

La matrice X est constituée de 20 lignes (20 échantillons) et de 2946 colonnes représentant des mesures d'absorbances entre 3950 et 1300 cm^{-1}. Le format de la matrice X est donc de 20×2946.

Le logiciel GRAM386-PLSPLUS permet de traiter ce type de données.

2.2.3. Etude exploratoire : analyse en composantes principales

Nous avons utilisé le logiciel GRAMS386-PLSPLUS pour réaliser l'analyse en composantes principales de la matrice X à 20 lignes et 2946 colonnes.

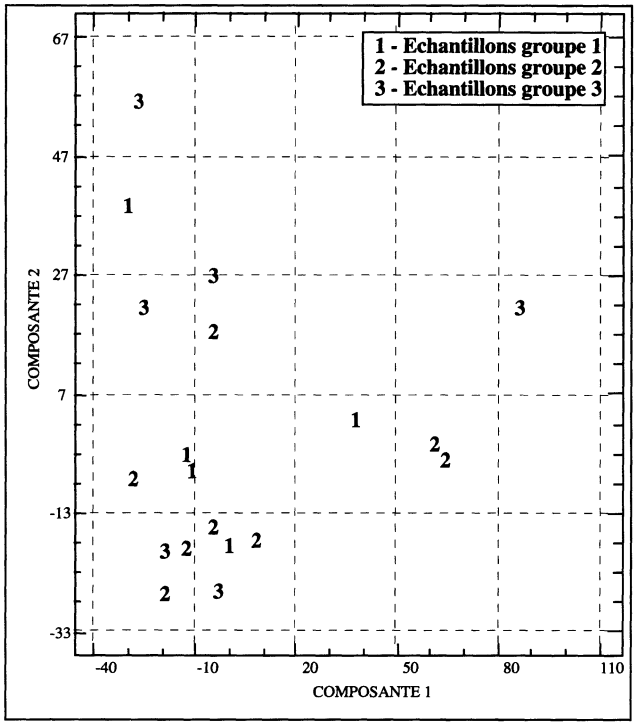


FIGURE 19
Analyse en composantes principales de la matrice X, composantes principales 1 et 2

Le premier plan principal (Figure 19) ne permet pas de séparer les groupes. Par contre on constate sur le graphique (CP1, CP3) de la figure 20 que la troisième composante principale apporte une meilleure discrimination des trois groupes initiaux.

2.2.4. Régression sur composantes principales

Nous avons réalisé une régression de chacune des variables Y_j sur les trois premières composantes principales du tableau X. Chaque échantillon est affecté au groupe pour lequel l'indicateur calculé $\widehat{Y}_k$ est le plus élevé. Les résultats sont présentés dans la figure 21.

On note 5 échantillons mal classés sur les 20 initiaux. En validation croisée on obtient 9 échantillons mal classés sur 20.

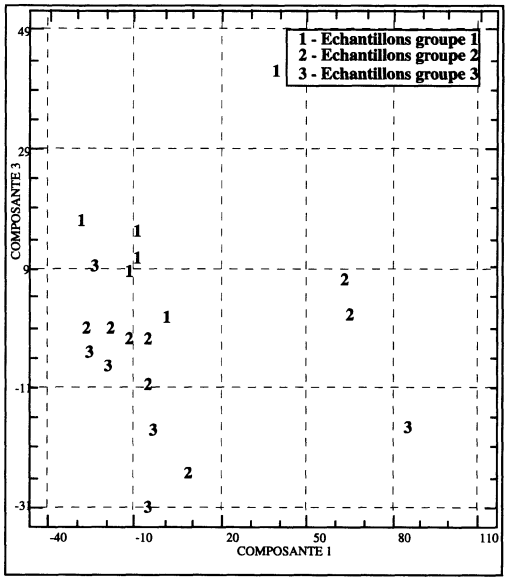


FIGURE 20
Analyse en composantes principales de la matrice X, composantes principales 1 et 3

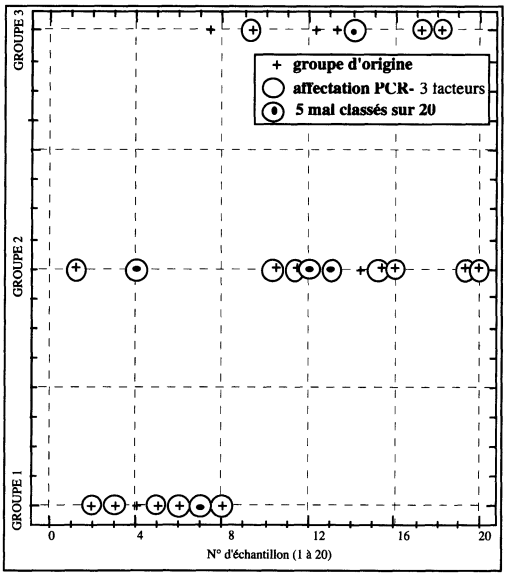


FIGURE 21
Régression sur composantes principales

2.2.5. La régression PLS

2.2.5.1. Régression PLS entre $Y = \{Y_1, Y_2, Y_3\}$ et le tableau X

Pour choisir le nombre de composantes PLS à retenir, on a utilisé le critère PRESS proposé dans GRAMS. Pour chaque échantillon on calcule l'erreur de prévision des Y_k sans utiliser cet échantillon pour construire le modèle. Le critère PRESS revient à choisir le nombre de composantes pour lequel la somme des carrés des erreurs de prévision est minimale. Ceci nous a conduit à retenir les trois premières composantes PLS t_1, t_2 et t_3.

Les plans (t_1, t_2) et (t_1, t_3) des figures 22 et 23 montrent que c'est l'axe 1 qui apporte la meilleure discrimination des trois groupes.

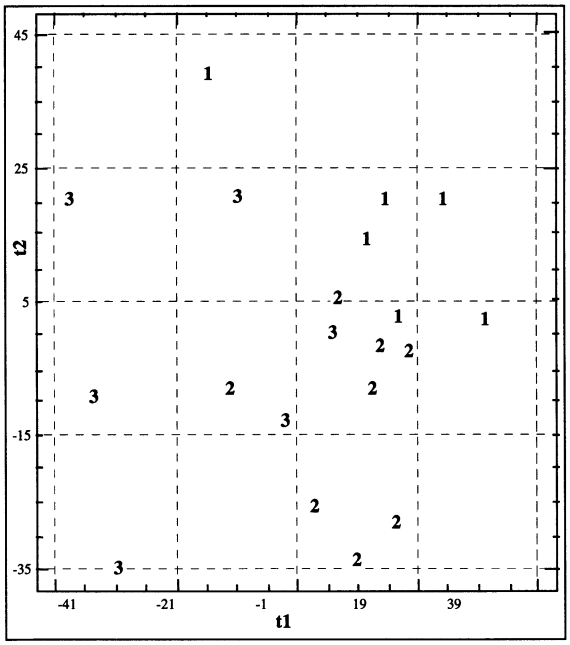


FIGURE 22
Régression PLS de Y sur X.
Représentation des échantillons dans le plan t_1, t_2

Dans la méthode de régression PLS, le calcul des axes est orienté vers l'explication de Y.

Utilisant des composantes t_h orientées vers l'explication de Y, la prévision des groupes d'origine va évidemment s'améliorer. Chaque échantillon est affecté au groupe pour lequel l'indicateur calculé $\widehat{Y}_k$ est le plus élevé. Les résultats sont présentés dans la figure 24.

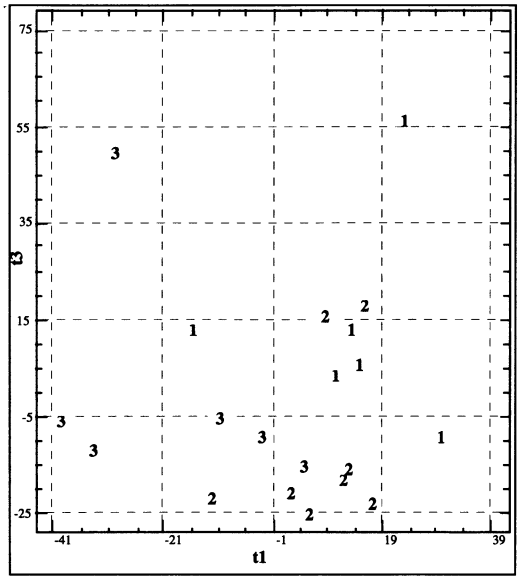


FIGURE 23
Régression PLS de Y sur X.
Représentation des échantillons dans le plan t_1, t_3

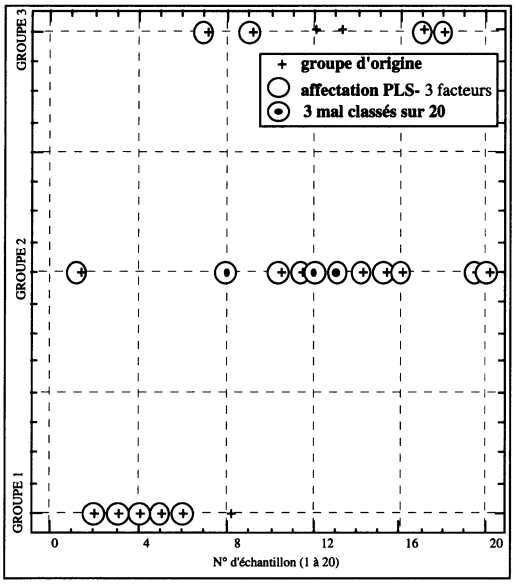


FIGURE 24
Régression PLS de Y sur X

On ne note maintenant que 3 échantillons mal classés sur les 20 initiaux. En validation croisée on obtient 7 échantillons mal classés sur 20.

2.2.5.2. Régression PLS de chaque Y_k sur X

On réalise maintenant trois régressions PLS indépendantes. Nous avons retenu les trois premières composantes PLS pour chaque régression PLS. On représente les trois plans (t_1, t_2) correspondant aux trois régressions PLS dans les figures 25, 26 et 27. On peut noter que l'axe 1 apporte bien la meilleure discrimination par rapport à chacun des trois groupes initiaux. On remarque aussi que dans les plans (t_1, t_2) de chaque analyse les groupes visés sont bien isolés.

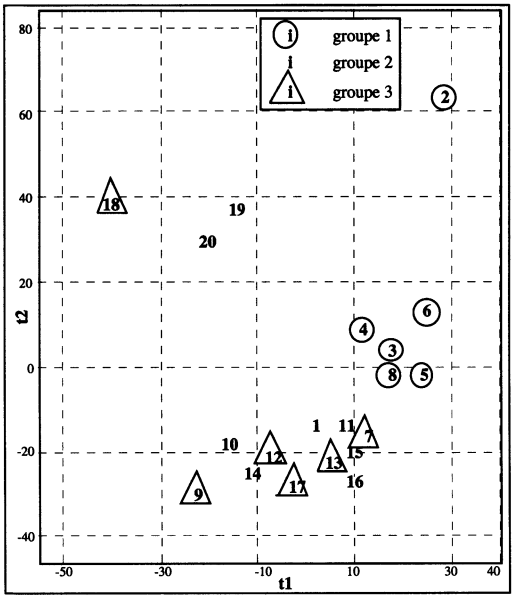


FIGURE 25
Régression PLS de Y_1 sur X

Chaque échantillon est affecté au groupe pour lequel l'indicateur calculé $\widehat{Y}_k$ est le plus élevé. Les résultats sont présentés dans la figure 28.

Il n'y a maintenant qu'un seul échantillon mal classé sur les 20 initiaux. En validation croisée le résultat s'améliore également : 5 mal classés sur 20.

2.2.6. Conclusion

Les travaux présentés dans cette partie mettent en évidence sur un cas réel l'intérêt de la régression PLS par rapport à la régression sur composantes principales. En régression PLS la méthode PLS-1 (on régresse chaque Y_k sur X) semble nettement plus performante que PLS-2 (on régresse Y sur X).

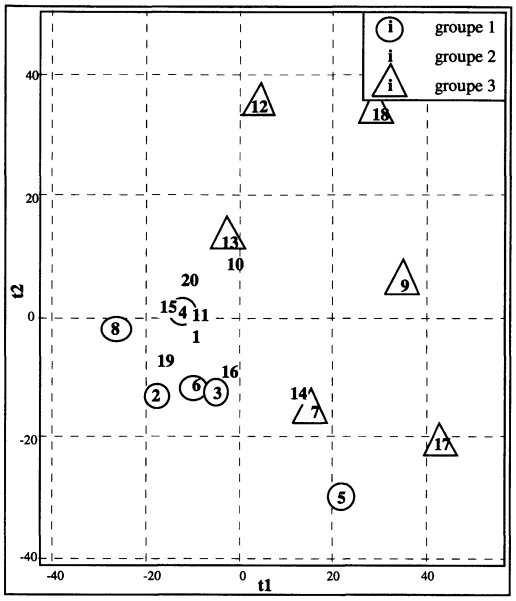


FIGURE 26
Régression PLS de Y_2 sur X

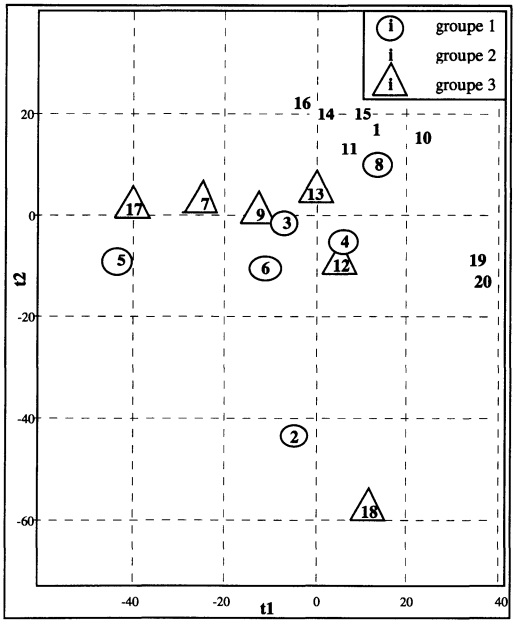


FIGURE 27
Régression PLS de Y_3 sur X

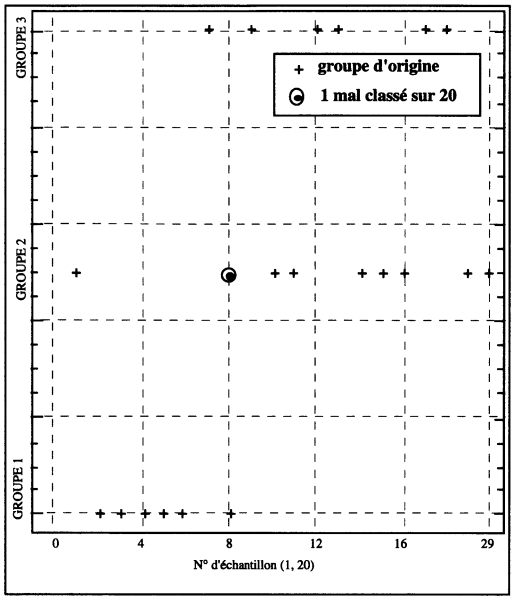


FIGURE 28
Régression PLS de chaque Y_k sur X

Cette étude de faisabilité permet d'envisager favorablement le remplacement d'un test long à mettre en œuvre par une analyse IR plus rapide.

2.3. La régression PLS qualitative

Il s'agit de relier un ensemble de variables dépendantes $Y = \{Y_1, \ldots, Y_K\}$ à un ensemble de variables indépendantes $X = \{X_1, \ldots, X_J\}$ lorsque ces variables sont qualitatives. Pour cela nous proposons la démarche suivante :

– On effectue une régression PLS des composantes principales issues de l'AFCM de Y sur les composantes principales issues de l'AFCM de X.

– On utilise ensuite la formule de reconstitution de données de l'AFCM pour prédire Y à partir des composantes principales des $Y_1, \ldots, Y_K$ estimées en fonction de X par régression PLS.

L'approche présentée est naturelle; elle peut aussi se justifier dans le cadre de la régression PLS généralisée à deux nuages de points appartenant à des espaces munis de métriques quelconques (Tenenhaus, 1993).

Cette procédure est résumée dans la figure 29.

Des applications pratiques de cette méthode sont en cours de réalisation.

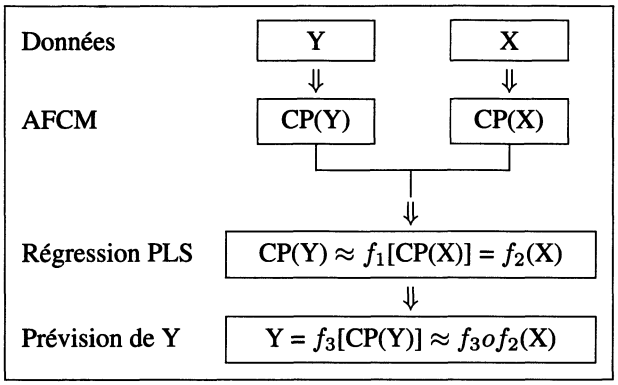


FIGURE 29
Régression PLS qualitative

Références Bibliographiques

CORNELL J.A. (1990). «Experiments with mixtures», Wiley.

FEDOROV V.V. (1972). «Theory of optimal experiments», translated (from russian, 1969) and edited by W.J. Studden and E.M. Klimbo, Academic Press, New York.

GAUCHI J.P. (1995). «Utilisation de la régression PLS pour l'analyse des plans d'expériences en chimie de formulation», *Revue de Statistique Appliquée*, 43, 1, p. 65-89.

GELADI P., KOWALSKI B.R. (1986). «Partial Least-Squares Regression : A Tutorial», *Anal. Chim. Acta*, 1, 185, p. 19-32.

GLEN W.G., DUNN III W.J., SCOTT D.R. (1989). «Principal Components Analysis and Partial Least Squares Regression», Tetrahedron Computer Methodology, 2, 6, p. 349-376.

GRAMS386 - PLSPLUS, Logiciel américain de la Société Galactic Industries Corporation Diffusé en France par la Société JMBS, Grenoble.

HÖSKULDSSON A. (1988). «PLS Regression Methods», J. Chemometrics, 2, p. 211-28.

JACKSON J.E. (1991). «A user's guide to principal components», Wiley.

KETTANEH-WOLD (1992). «Analysis of mixture data with partial least squares», Chemometrics and Intelligent Laboratory Systems, 14, p. 57-69.

KVALHEIM O.M. (1988). «A Partial-Least Squares Approach to Interpretative Analysis of Multivariate Data». Chemometrics and Intelligent Laboratory Systems, 3, p. 189-197.

MARTENS H., NAES T. (1989). «Multivariate Calibration», Wiley.

MENARDO C. (1993). «Application de la régression PLS à la spectrométrie IRTF», *Revue de Modélisation et Analyse des Données*, n°5, p. 37-46.

SIMCA (1991). «Soft Independent Modeling of Class Analogy», Version 4.3R, Umetri AB Box 1456, S-901 24 Umea.

TENENHAUS M. (1993). «La régression PLS généralisée», *Cahier de recherche* n° 472, Groupe HEC, Jouy-en-Josas.

WOLD S. (1989). «Multivariate Data Analysis : Converting Chemical data Tables to Plots», in : Computer Applications in Chemical Research and Education. Dr. Alfred Hüthig Verlag, Heidelberg.