1.2. Régression PLS (Partial Least Squares regression) [6, 7, 16, 18, 19]

La régression PLS (Partial Least Squares regression) est une méthode d'analyse des données proposée par Wold, Albano, Dunn III, Esbensen, Hellberg, Johansson & Sjöström (1983). Cette méthode connaît un très grand succes dans le domaine de la chimie, particulièrement dans les applications concernant des données de chromatographie ou de spectrographie. Elle s'est principalement développée autour de Svante Wold, Professeur au Department of Organic Chemistry, Research Group for Chemometrics, University of Umeå, Umeå, Sweden. De plus Svante Wold, Nouna Kettanech-Wold et leurs collaborateurs ont développé le logiciel d'analyse des données SIMCA-P for Windows centré sur la régression PLS. Signalons également l'aventage de la régression PLS par rapport à d'autres méthodes de régression dans l'analyse des plans d'expériences non orthogonaux *(Gauchi, 1995)*

La régression PLS devrait pouvoir s'appliquer à de nombreux domaines avec le même succès qu'en chimie.

Au niveau pratique, la régression PLS existe dans les logiciels SIMCA-P for Windows (Umetrics AB) et The Unscrambler (camo AS) aussi que plusieurs macros pour excel. Le logiciel SIMCA-P ayant été développé par le fondateur de la méthode. Par conséquent nous avons tenu à ajuster les résultats de la régression PLS au sorties du logiciel SIMCA-P+. Précisons cependant qu'une certaine standardisation existe : les sorties du logiciel The Unscrambler et de la Proc PLS de SAS sont très voisines de celles de SIMCA-P.

Historiquement la régression PLS est née de l'algorithme NIPALS développé par H. Wold (1966). Cet algorithme est essentiel à la compréhension de la régression PLS car il permet de comprendre la logique de l'écriture de l'algorithme de régression PLS. On distingue habituellement en régression PLS le cas où il y a une seule variable Y de celui où il y en a plusieurs. La régression PLS1 correspond à la première situation et la régression PLS2 à la seconde. L'algorithme de régression PLS a été proposé initialement par Wold, Martens & Wold (1983) et Wold, Albano, Dunn III, Esbensen, Hellberg, Johanssion & Sjöström (1983).

1.2.1 Principe des régressions RCP et PLS [16, 18]

Une alternative intéressante à la régression linéaire multiple consiste à remplacer une matrice des données prédictives ***X*** comprenant ***n*** lignes et ***m*** colonnes, par une nouvelle matrice, dérivée de ***X***, qu'on design par ***T***, comprenant le même nombre de lignes (observations) que ***X*,** mais un nombre de colonnes ***k*** très inférieur à ***m*.** On impose, de plus, que les colonnes de la matrice ***T*** soient des combinaisons linéaires des variables d'origine.

Sous forme matricielle, la relation s'écrire :

$$\boldsymbol{T = XW} \tag{1.5}$$

Avec

W la matrice de dimensions ***m*** × ***k*** des coefficients définissant les combinaisons linéaires.

T est donc une nouvelle matrice dont les colonnes forment des « variables artificielles », obtenues par combinaison linéaire des variables d'origine.

Après cette transformation, la régression linéaire multiple est appliquée sur le tableau ***T*** à la place de ***X***. Le problème est donc de déterminer ***W*** de manière à avoir une matrice de variables prédictives ***T*** plus adaptée au calcul de la régression que la matrice ***X*** d'origine.

Deux méthodes de régression : la **régression sur composantes principales (RCP)** et la **régression PLS** sont fondées sur cette approche. La différence principale entre ces deux méthodes réside dans leur manière de calculer la matrice ***W***.

1.2.2 Régression sur composantes principales (RCP)

L'**analyse en composantes principales** (ACP) est une méthode statistique qui effectue précisément l'opération recherchée.

À partir d'un tableau de données ***X*** centré, l'ACP donne, comme résultats principaux, la matrice ***T*** dont les colonnes (appelées *composantes principales*) sont orthogonales entre elles et la matrice ***W*** des vecteurs définissant les coefficients des combinaisons linéaires ou matrice des **vecteurs propres** de $\boldsymbol{X}^T\boldsymbol{X}$. De plus, les colonnes de ***T*** sont également caractérisées par leur variance, qui sont également les **valeurs propres** de $\boldsymbol{X}^T\boldsymbol{X}$.

Lorsque les données ***X*** sont colinéaires, le nombre ***k*** de colonnes de ***T*** est nécessairement inférieur ou égal à ***m***, le nombre de colonnes de ***X***. Le nombre ***k*** désigne le *rang* de la matrice ***X***. Il correspond également au nombre de valeurs propres non nulles.

À partir de ***T***, il est, en principe, aisé d'établir un modèle de régression de la forme :

$$\hat{y}_i = q_1 t_{i1} + q_2 t_{i2} + \ldots + q_k t_{ik} \quad \text{pour } i = 1, \ldots, n \qquad (1.6)$$

Avec, t_{ij}, élément de ***T*** en ligne ***i*** et en colonne ***j***.

1.2.3 Régression PLS (*Partial Least-Squares*) [7, 18, 19]

La régression PLS est la méthode de régression **la plus couramment utilisée** dans les analyses spectrométriques. Elle est bien implantée dans de nombreux logiciels commerciaux, sous une forme conviviale comme SIMCA-P+ que nous avons utilisé pendant le stage.

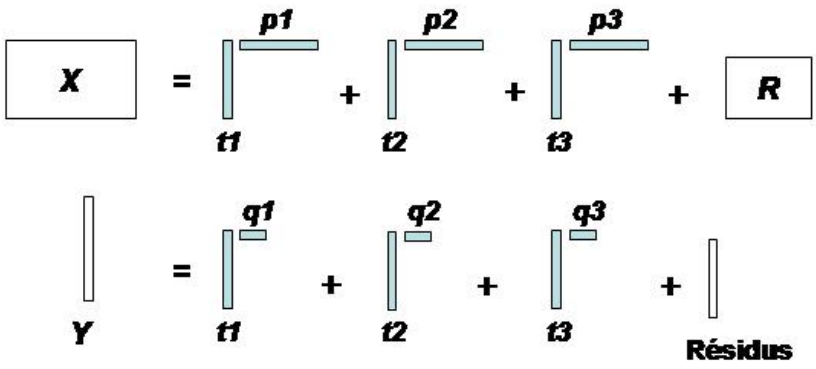

Figure 1.4 - Modèle de régression PLS (exemple d'un modèle à 3 dimensions)

1.2.3.1. Établissement du modèle PLS simple [18]

La régression PLS présente de nombreux points communs avec la RCP. Elle comprend également une étape de calcul d'une matrice ***T***, dont les éléments sont les *scores* et les colonnes les *composantes,* obtenue par l'application d'une relation de la forme :

$$\boldsymbol{T = XW} \qquad (1.7)$$

Avec

W matrice des poids ou *loadings*

X matrice des données prédictives centrées, comprenant ***n*** lignes et ***m*** colonnes.

Dans la régression PLS, le calcul des composantes ***T*** se fait en tenant compte de la variable à prédire ***y***. Plus précisément, on cherche à effectuer une double modélisation (figure 1.4) correspondant aux deux relations :

$$\boldsymbol{X = TP + R} \qquad (1.8)$$

$$\boldsymbol{y = Tq + f} \qquad (1.9)$$

Avec

R, matrice des résidus, associée à la prédiction de ***X***

F, vecteur des résidus associé à la prédiction de ***y***.

La première étape consiste à calculer $\boldsymbol{t_1}$, la première composante. On estime ensuite les paramètres des modèles (1.8) et (1.9) à une seule composante, soit :

$$\boldsymbol{X = t_1 p_1 + R_1} \text{ et } \boldsymbol{y = t_1 q_1 + f_1} \qquad (1.10)$$

Avec

$\boldsymbol{t_1}$, de dimension $n \times 1$

$\boldsymbol{p_1}$, de dimension $1 \times m$

$\boldsymbol{q_1}$, de dimension 1×1 (nombre).

Le deuxième modèle associé à ***X*** ainsi obtenu indique qu'une *ligne* $\boldsymbol{x_i}$ d'indice ***i*** de ***X*** est égale à la somme de deux « signaux purs » représentés par les vecteurs $\boldsymbol{p_1}$ et $\boldsymbol{p_2}$ pondérés par les valeurs $\boldsymbol{t_{1i}}$ et $\boldsymbol{t_{2i}}$, soit :

$$\boldsymbol{x = t_{1i}\ p_1 + t_{2i}\ p_2} \qquad (1.11)$$

On voit aisément que $\boldsymbol{p_1}$ et $\boldsymbol{p_2}$ ont une signification claire, et peuvent se présenter graphiquement d'une manière homologue à un spectre.

L'introduction de nouvelles composantes dans le modèle se fait selon la même procédure : ayant un modèle à ***k*** composantes (ou ***k*** *dimensions*), on crée un nouveau modèle à $k + 1$ dimensions en calculant tout d'abord une nouvelle composante t_k, puis les paramètres des deux modèles couplés :

$$X = t_1 p_1 + t_2 p_2 + ... + t_k p_k + t_{k+1} p_{k+1} + R_{k+1} \quad (1.12)$$

$$y = t_1 q_1 + t_2 q_2 + ... + t_k q_k + t_{k+1} q_{k+1} + f_{k+1} \quad (1.13)$$

Ce qui s'écrit, de manière équivalente, sous les formes matricielles (1.12) et (1.13).

À une étape ***k*** donnée, la composante t_k est déterminée à partir des résidus R_{k-1} et f_{k-1} de l'étape précédente.

1.2.3.3 La régression PLS2 [19]

La régression PLS2 s'applique dans le cas où plusieurs variables sont à prédire à partir de la même matrice de données prédictives ***X***. Les données à prédire forment maintenant la matrice ***Y***, comprenant ***n*** lignes et ***r*** colonnes.

La démarche et les algorithmes sont très voisins de ceux suivis dans la régression PLS simple (généralement dénommée *PLS1*) décrits précédemment.

Considérons les données initiales centrées X_0 et Y_0. On cherche à calculer une première composante t_1, combinaison linéaire des colonnes de X_0 qui soit à la fois représentative de X_0 et de Y_0. Pour cela, on s'intéresse tout d'abord à deux combinaisons linéaires associées respectivement à X_0 et à Y_0 soit :

$$t_1 = X_0 w_1 \text{ et } u_1 = Y_0 c_1 \quad (1.14)$$

Avec w_1 et c_1 vecteurs de norme unitaire (à déterminer).

Le critère à optimiser est une extension, dans le cas où ***Y*** est multivariée, de celui adopté pour la régression PLS1. On détermine w_1 et c_1 de manière à **maximiser la covariance** entre t_1 et u_1. Ce critère conduit à résoudre un système d'équations aux valeurs propres que nous ne détaillerons pas ici.

L'algorithme est ensuite presque identique à celui de la régression PLS1. On effectue à nouveau une **double modélisation** :

$$X_0 = t_1 p_1 + X_1 \text{ et } Y_0 = t_1 q_1 + Y_1 \quad (1.15)$$

Les éléments de $\boldsymbol{q_1}$ sont les coefficients des régressions linéaires simples effectuées entre $\boldsymbol{t_1}$ et chaque colonne de $\boldsymbol{Y_0}$. Dans cette étape, la nature symétrique de l'algorithme disparaît puisque à la fois $\boldsymbol{X_0}$ et $\boldsymbol{Y_0}$ sont estimés à partir de $\boldsymbol{t_1}$, qui est une combinaison linéaire des colonnes de $\boldsymbol{X_0}$, tandis que $\boldsymbol{u_1}$, combinaison linéaire des colonnes de $\boldsymbol{Y_0}$, n'intervient pas dans ces relations.

Les itérations suivantes sont similaires à cette première étape et sont également analogues à celles de PLS1. On peut, comme précédemment, estimer les coefficients de la régression PLS s'appliquant directement sur la matrice $\boldsymbol{X}$. Pour chaque dimension testée notée $\boldsymbol{k}$, ces coefficients peuvent être regroupés dans une matrice $\boldsymbol{B_k}$ de dimensions $\boldsymbol{m} \times \boldsymbol{r}$.