La régression Partial Least-Squares boostée

Jean-François DURAND[1]

Université Montpelier II, France
E-Mail : jf.durand@club-internet.fr
Site web : www.jf-durand-pls.com

Résumé Ce papier présente la régression Partial Least-Squares (PLS) comme appartenant à la famille de méthodes de boosting à fonction coût L_2. D'une part, la régression PLS linéaire classique appartient à cette catégorie en considérant une variable latente ou composante principale comme base d'apprentissage (*base learner*) rendant robuste le modèle face au problème de la rareté des données et de la multi-corrélation des variables. D'autre part, l'usage des B-splines et de leurs produits tensoriels dans la construction de la base d'apprentissage, exploite de façon naturelle le potentiel du boosting L_2 de PLS pour produire des modèles non-linéaires additifs qui capturent les effets principaux ainsi que les interactions significatives. La performance du boosting PLS en régression comme en classification supervisée est montrée sur trois exemples.
Keywords :Boosting, Partial Least-Squares, B-splines, Produits Tensoriels
This paper presents the Partial Least-Squares regression (PLS) in the framework of the boosting methods with L_2 loss. First, the ordinary PLS regression already belongs to that family by considering the latent variables or principal components as base learners producing robust linear models that overcome the problems of the scarcity of the observations as well as the multi-collinearity of the predictor variables. Most of all, the use B-splines and their tensor products to construct the base learner, typically provides PLS with L_2-boosts leading to non-linear additive models that capture main effects as well as relevant interactions. The performances of the different PLS boosts in both regression and classification are shown on three exemples.
Keywords : Boosting, Partial Least-Squares, B-splines, Tensor Products

1 Introduction

La régression linéaire Partial Least-Squares (Wold et al., 1983), (Tenenhaus, 1998), en bref PLS, est une méthode statistique de prévision très populaire en chimiométrie à sa création et maintenant dans tous les domaines industriels et économiques. Les débuts de PLS dans le monde académique ont été plus laborieux peut-être à cause de sa formulation algorithmique. Depuis quelques années cependant, la recherche statistique a pris à son compte l'introduction d'algorithmes dans la définition des méthodes de prévision. Un exemple est donné par le boosting L_2 dont la version séminale est le "twicing" (Tukey, 1977). L'idée consiste à améliorer une méthode peu performante comme celle des moindres carrés, en l'appliquant de façon récurrente sur les données mal prédites que sont les résidus, pour construire, étape par étape, un modèle additif. Les méthodes modernes de boosting transforment à chaque étape, les variables explicatives par une fonction de classe paramétrique, appelée la base d'apprentissage (*base learner*). Cela peut être un arbre de décision (Hastie et al., 2001), une spline de lissage (Bühlmann et Bin Yu, 2000)...

Le premier objectif de cet article est de montrer comment une variable latente PLS, combi-

naison linéaire des variables explicatives et de covariance maximale avec les réponses, peut être considérée comme base d'apprentissage situant ainsi la régression PLS comme une des premières de l'histoire du boosting L_2. C'est une des explications de l'efficacité de PLS pour capturer des relations linéaires dans le difficile contexte de la rareté des observations d'apprentissage et de la multi-colinéarité entre variables explicatives. Cependant, le boosting PLS exploite mieux son potentiel en enrichissant la base d'apprentissage par la transformation de chaque variable explicative par une base de B-splines. Une variable latente devient alors une somme de splines dîtes "fonctions coordonnées" car dépendant chacune d'une seule variable. Le modèle additif obtenu (Durand, 2001), appelé Partial Least-Squares Splines, PLSS en bref, est une alternative robuste à la régression Least-Squares Splines (Stone, 1985) face au sur-ajustement des données, voir (Avelino et al., 2007) pour une application récente de PLSS à l'étude du développement du *mycena citricolor*, maladie du caféier, au Costa Rica. Si PLSS produit un modèle purement additif capturant les effets principaux et non les interactions, l'usage de produits tensoriels de B-splines est bien adapté aux interactions lorsqu'il s'accompagne d'une procédure pour éliminer celles qui sont négligeables. C'est la caractéristique du modèle Multivariate Additive Partial Least-Squares Splines (Durand et Lombardo, 2003), (Lombardo et al., soumis), en bref MAPLSS, qui se présente comme une décomposition de type ANOVA des effets principaux et des éventuelles interactions bi-variées les plus significatives.

Le prix à payer par ces méthodes pour la capture des non-linéarités et des interactions, est l'expansion de la dimension qui est traditionnellement bien supportée par PLS. L'autre objectif de cet article est de faire le point sur ces récents développements, sur le rôle des super-paramètres et sur la pratique de la construction des modèles. Le site internet *www.jf-durand-pls.com* propose le logiciel libre dédié au boosting PLS, appelé *PLSS*, et écrit sous *R* (R development core team, 2006). Il contient les sources et le descriptif des fonctions utilisées dans cet article ainsi que les transparents sur la régression PLS boostée du cycle J.P. Fénelon.

La Section 2 présente le boosting L_2 sous sa forme d'algorithme fonctionnel de descente de type gradient. La définition et le rôle de la base d'apprentissage sont précisés. Une légère généralisation de l'algorithme est proposée ne modifiant en rien la philosophie de la méthode qui ajuste à chaque étape les résidus aux moindres carrés pour construire pas-à-pas, un modèle linéaire en la base d'apprentissage. La régression PLS ordinaire, résumée dans la Section 3, est alors présentée comme une méthode de boosting L_2 avec pour bases d'apprentissage les variable latentes ou composantes principales. Les bases de B-splines sont introduites dans la Section 4 qui expose PLSS et MAPLSS, les deux extensions du modèle linéaire pour la capture des non-linéarités et des interactions. La Section 5 donne des aides à la construction des modèles par des choix raisonnés des paramètres splines. Trois exemples sur des données réelles et simulées sont présentés Section 6, pour illustrer la mise en application et la performance des méthodes dans les deux contextes de la régression et de la classification.

2 Le boosting, un algorithme fonctionnel de descente de type gradient

Un pas important dans la justification théorique du bon comportement du boosting a été franchi en le présentant comme un algorithme global d'optimisation (Breiman, 1999), (Hastie et al., 2001), (Friedman, 2001). Notons y la variable réponse qui peut être continue, problème de régression, ou discrète, problème de classification et $\boldsymbol{x} \in \mathbb{R}^p$ la variable explicative. Il s'agit d'estimer, à partir des données $(\boldsymbol{x}_i, y_i)_{i=1,\ldots n}$, la fonction $F : \mathbb{R}^p \to \mathbb{R}$, minimisant l'espérance

du coût

$$I\!E[L(y, F(\boldsymbol{x})], \qquad L(.,.) : I\!R \times I\!R \to I\!R^+. \tag{1}$$

La fonction coût $L(.,.)$ peut prendre différentes expressions suivant le type de prolème à traiter et de nombreuse variantes sont étudiées dans (Hastie et al., 2001). Pour rester au plus près de l'objectif fixé, seul le coût L_2, $L(y, f) = \frac{1}{2}(y - f)^2$, est ici envisagé. Un minimiseur de (1) est de la forme $F(x) = I\!E[y|\boldsymbol{x} = x]$ dont l'estimation est contrainte d'une part, à être de type additif

$$\hat{F}(\boldsymbol{x}\,;\,\{\alpha_m, \boldsymbol{\theta}_m\}_1^M) = \sum_{m=1}^{M} \alpha_m h(\boldsymbol{x}, \boldsymbol{\theta}_m), \tag{2}$$

et d'autre part, à dépendre d'une fonction paramétrique, $h(\boldsymbol{x}, \boldsymbol{\theta}) : I\!R^p \times \Theta \to I\!R$, appelée la base d'apprentissage. La base d'apprentissage a pour objectif la capture de non-linéarités et d'interactions et puise ses exemples parmi les arbres de décision, les splines, les ondelettes...

Dans la Section 3, nous verrons que pour PLS ordinaire, $\Theta = I\!R^p$ et que la base d'apprentissage est une variable latente ou composante principale,

$$t = h(\boldsymbol{x}, \boldsymbol{\theta}) =< \boldsymbol{x}, \boldsymbol{\theta} >, \tag{3}$$

où $< .,. >$ est le produit scalaire ususel de $I\!R^p$. Lorsque Θ est un espace vectoriel de dimension finie, la dimension de la base d'apprentissage est $r = dim\,\Theta$. Section 4, les bases d'apprentissage PLS non-linéaires sont du même type avec $\Theta = I\!R^r$, $r \geq p$. Le coût àà payer pour la capture des non-linéarités et interactions est bien supporté par PLS qui est robuste face au problème de la dimension $r \gg n$.

Le modèle (2) issu du boosting est construit pas-à-pas en M étapes, chacune prenant en compte la partie mal prédite par la précédente. La dimension du modèle, M, est généralement estimée par validation-croisée. Le boosting est un algorithme de descente de type gradient calculé par rapport à F.

Algorithme 1 (boosting L_2).

1. $F_0 = 0$

2. For $m = 1$ to M do

3. $\quad \tilde{y}_i = -\left[\frac{\partial L(y_i,F)}{\partial F}\right]_{F=F_{m-1}(\boldsymbol{x}_i)} = y_i - F_{m-1}(\boldsymbol{x}_i), \quad i = 1, \ldots, n,$

4. $\quad (\alpha_m, \boldsymbol{\theta}_m) = \arg\min_{\alpha,\boldsymbol{\theta}} \sum_{i=1}^{n} [\tilde{y}_i - \alpha h(\boldsymbol{x}_i; \boldsymbol{\theta})]^2$

5. $\quad F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \alpha_m h(\boldsymbol{x}; \boldsymbol{\theta}_m)$

6. endFor

Ligne 3 de l'Algorithme 1, la pseudo-réponse $\tilde{y}$, gradient par rapport à F de la fonction coût, est le résidu courant. Ligne 4, le paramètre de la base d'apprentissage est estimé par les moindres carrés. Ce n'est pas le cas pour PLS où une variable latente est estimée par un critère de covariance. Pour assimiler PLS aux méthodes de boosting, la légère généralisation suivante décompose la phase 4 en deux parties, l'une pour estimer $\boldsymbol{\theta}_m$, l'autre pour estimer le paramètre α_m.

Algoritme 2 (boosting L_2 étendu) : Dans l'Algorithme 1, remplacer 4. par

4.1 Critère de proximité entre $\boldsymbol{x}$ et y pour construire $\boldsymbol{\theta}_m$ à partir des données,
4.2 $\alpha_m = \arg\min_\alpha \sum_{i=1}^n [\tilde{y}_i - \alpha h(\boldsymbol{x}_i; \boldsymbol{\theta}_m)]^2$.

L'Algorithme 2 est identique à l'Algorithme 1 lorsque, ligne 4.1, les moindres carrés, (Bühlmann et Bin Yu, 2000), sont utilisés pour estimer la base d'apprentissage.

En résumé, une méthode de boosting L_2 se caractérise par l'ajustement à chaque étape des résidus courants et l'élaboration d'un modèle additif (2) construit pas-à-pas par une combinaison linéaire des bases d'apprentissage. La dimension du modèle, M, est le nombre de bases d'apprentissage utilisées.

3 PLS, un algorithme de boosting L_2

Il est habituel d'introduire la régression PLS par des considérations géométriques basées sur des notations matricielles. Les données d'apprentissage sont notées $(\boldsymbol{X}, \boldsymbol{Y})$, où $\boldsymbol{X} = [x_i^j]$ est la matrice $n \times p$ des observations sur les p variables explicatives, $\boldsymbol{Y} = [y_i^j]$, $n \times q$, celle des observations sur les q variables réponses. PLS est ici présentée comme une méthode de régression multi-réponses, un exemple typique est celui où les réponses sont les indicatrices binaires de groupes d'individus. Dans ce cas PLS est une méthode décisionnelle pour diagnostiquer l'appartenance d'un individu à un groupe. Les colonnes des matrices $\boldsymbol{X}$ et $\boldsymbol{Y}$ sont centrées (généralement réduites) par rapport à la matrice diagonale $\boldsymbol{D} = diag(p_1, \ldots, p_n) = n^{-1}\boldsymbol{I}_n$ des poids statistiques des observations. Ainsi, la covariance empirique entre deux variables (colonnes) devient le $\boldsymbol{D}$-produit scalaire de $\mathbb{R}^n$

$$cov(\boldsymbol{x}^j, \boldsymbol{y}^k) = < \boldsymbol{x}^j, \boldsymbol{y}^k >_{\boldsymbol{D}} = \boldsymbol{y}^{k\prime}\boldsymbol{D}\boldsymbol{x}^j \text{ et } var(\boldsymbol{x}^j) = \|\boldsymbol{x}^j\|^2_{\boldsymbol{D}}.$$

Table 1 : L'algorithme PLS vu comme un algorithme de boosting L_2.

Initialisation	$\boldsymbol{X}_{(0)} = \boldsymbol{X}, \quad \boldsymbol{Y}_{(0)} = \boldsymbol{Y}$	
Étape m	4.1 La base d'apprentissage	$\boldsymbol{t} = \boldsymbol{X}_{(m-1)}\boldsymbol{w}$, $\boldsymbol{u} = \boldsymbol{Y}\boldsymbol{v}$
	4.1.1 Construction de	$(\boldsymbol{w}^m, \boldsymbol{v}^m) = \arg \max_{\boldsymbol{w}'\boldsymbol{w}=1=\boldsymbol{v}'\boldsymbol{v}} cov(\boldsymbol{t}, \boldsymbol{u})$
	$\boldsymbol{t}^m \in Im\, \boldsymbol{X}_{(m-1)}$	$\boldsymbol{t}^m = \boldsymbol{X}_{(m-1)}\boldsymbol{w}^m$, $\boldsymbol{u}^m = \boldsymbol{Y}\boldsymbol{v}^m$
	.....	.....
m=1,...,M	4.1.2 Déflation de $\boldsymbol{X}_{(m-1)}$	$\boldsymbol{X}_{(m)} = \boldsymbol{X}_{(m-1)} - \boldsymbol{H}_m \boldsymbol{X}_{(m-1)}$
	4.2. Résidus de $\boldsymbol{Y}_{(m-1)}$	$\boldsymbol{Y}_{(m)} = \boldsymbol{Y}_{(m-1)} - \boldsymbol{H}_m \boldsymbol{Y}_{(m-1)}$

Table 1, l'algorithme PLS construit , étape par étape, des variables latentes ou composantes principales, notées $\{\boldsymbol{t}^m\}_{m=1,\ldots,M}$, compromis linéaires des pseudo variables explicatives $\boldsymbol{X}_{(m-1)}$, mais aussi, avec (5), des variables naturelles $\boldsymbol{X}$, et de covariance maximale avec des combinaisons linéaires des réponses. L'originalité de PLS tient dans le fait que l'étape courante est basée sur les pseudo variables explicatives, $\boldsymbol{X}_{(m-1)}$, et non directement sur les variables initiales $\boldsymbol{X}$, sauf à l'étape $m = 1$ où $\boldsymbol{X}_{(0)} = \boldsymbol{X}$. La phase 4.1.2 est l'actualisation des pseudo-explicatives

par la déflation de la matrice $\boldsymbol{X}_{(m-1)}$. Ligne 4.2, la régression aux moindres carrés, $\boldsymbol{H}_m \boldsymbol{Y}_{(m-1)}$, dîte partielle, des pseudo-réponses $\boldsymbol{Y}_{(m-1)}$ sur la composante permet leur actualisation en tant que résidus courants. La matrice du régresseur, $n \times n$ et de rang 1,

$$\boldsymbol{H}_m = \boldsymbol{t}^m \boldsymbol{t}^{m\prime} \boldsymbol{D} / \|\boldsymbol{t}^m\|^2_{\boldsymbol{D}} \tag{4}$$

est le projecteur $\boldsymbol{D}$-orthogonal sur $\boldsymbol{t}^m$. Nous verrons dans (8) comment le modèle (2) se construit pas-à-pas et dans (11), devient linéaire en les variables $\boldsymbol{X}$.

La phase 4.1.1 de construction de $\boldsymbol{t}^m$, est un problème d'optimisation à deux contraintes portant sur les poids $\boldsymbol{w}$ et $\boldsymbol{v}$ assujettis à être de norme 1. La solution obtenue par la technique des multiplicateurs de Lagrange, est donnée par le premier terme de la décomposition en valeurs singulières de la matrice $p \times q$ des covariances $\boldsymbol{X}'_{(m-1)} \boldsymbol{D} \boldsymbol{Y}$. Soit $(\lambda_m, \boldsymbol{w}^m, \boldsymbol{v}^m)$ le triplet correspondant à la plus grande valeur singulière λ_m, alors

$$\boldsymbol{t}^m = \boldsymbol{X}_{(m-1)} \boldsymbol{w}^m, \qquad \boldsymbol{u}^m = \boldsymbol{Y} \boldsymbol{v}^m, \qquad \lambda_m = cov(\boldsymbol{t}^m, \boldsymbol{u}^m).$$

Dans le cas d'une seule réponse, $\boldsymbol{v}^m = 1$ et $\boldsymbol{w}^m = \boldsymbol{X}'_{(m-1)} \boldsymbol{D} \boldsymbol{Y} / ||\boldsymbol{X}'_{(m-1)} \boldsymbol{D} \boldsymbol{Y}||$.

Trois propriétés cruciales des composantes $\{\boldsymbol{t}^m\}$:

– Elles sont dans $Im\ \boldsymbol{X}$, i.e.,

$$\exists\, \boldsymbol{\theta}^m = (\theta^m_1, \ldots, \theta^m_p)' \qquad \text{tel que} \qquad \boldsymbol{t}^m = \boldsymbol{X} \boldsymbol{\theta}^m. \tag{5}$$

La construction de $\boldsymbol{\theta}^m$ met un point final à la phase 4.1 de l'algorithme du boosting L_2. Cette construction s'effectue pas-à-pas avec au départ, $\boldsymbol{\theta}^1 = \boldsymbol{w}^1$. Pour plus de détails voir, par exemple, le livre de M. Tenenhaus (1998), page 107, où $\boldsymbol{\theta}$ est noté $\boldsymbol{w}^*$.

– Elles sont mutuellement $\boldsymbol{D}$-orthogonales, $\boldsymbol{t}^{m_1\prime} \boldsymbol{D} \boldsymbol{t}^{m_2} = 0, \quad m_1 \neq m_2$.

– Les régressions partielles des pseudo variables donnent les mêmes résultats que les régressions aux moindres carrés des variables originelles sur les composantes. Ceci permet de construire

$$\boldsymbol{p}^m \in I\!R^p, \text{ défini par } \boldsymbol{H}_m \boldsymbol{X}_{(m-1)} = \boldsymbol{H}_m \boldsymbol{X} = \boldsymbol{t}^m \boldsymbol{p}^{m\prime},$$

$$\boldsymbol{\alpha}^m \in I\!R^q, \text{ défini par } \boldsymbol{H}_m \boldsymbol{Y}_{(m-1)} = \boldsymbol{H}_m \boldsymbol{Y} = \boldsymbol{t}^m \boldsymbol{\alpha}^{m\prime}. \tag{6}$$

La régression partielle $\boldsymbol{H}_m \boldsymbol{Y}$ termine l'étape 4 de l'algorithme du boosting L_2 et (6) exhibe le coefficient $\boldsymbol{\alpha}^m = (\alpha^m_1, \ldots, \alpha^m_q)'$.

Le modèle PLS linéaire en les bases d'apprentissage $\{\boldsymbol{t}^m\}_1^M$

L'écriture successive des M équations 4.1.2 et 4.2 de l'algorithme PLS conduit à l'expression de l'ajustement tant du côté des variables explicatives que du côté des réponses. En effet, PLS produit grâce à $\hat{\boldsymbol{Y}}(M)$ un ajustement des réponses $\boldsymbol{Y}$, tout en tentant, étape par étape, de reconstruire par $\hat{\boldsymbol{X}}(M)$, la matrice $\boldsymbol{X}$ des variables explicatives

$$\hat{\boldsymbol{X}}(M) = \boldsymbol{t}^1 \boldsymbol{p}^{1\prime} + \ldots + \boldsymbol{t}^M \boldsymbol{p}^{M\prime}, \qquad \text{et} \qquad \boldsymbol{X} = \hat{\boldsymbol{X}}(M) + \boldsymbol{X}_{(M)} \tag{7}$$

$$\hat{\boldsymbol{Y}}(M) = \boldsymbol{t}^1 \boldsymbol{\alpha}^{1\prime} + \ldots + \boldsymbol{t}^M \boldsymbol{\alpha}^{M\prime}, \qquad \text{et} \qquad \boldsymbol{Y} = \hat{\boldsymbol{Y}}(M) + \boldsymbol{Y}_{(M)}\,. \tag{8}$$

Dans le modèle (8), $\hat{\boldsymbol{Y}}(M)$ est la version PLS, matricielle et multi-réponses, de l'ajustement (2) du boosting L_2.

Si $M = rang(\boldsymbol{X})$, alors, $\{\boldsymbol{t}^1, \ldots, \boldsymbol{t}^M\}$ forment une base de $Im\,\boldsymbol{X}$, et $\boldsymbol{X}_{(M)} = \boldsymbol{0}$. Dans ce cas, PLS reconstruit exactement $\boldsymbol{X}$ et (8) donne les modèles aux moindres carrés ordinaires. Notons de façon abrégée,

$$\text{si} \qquad M = rang(\boldsymbol{X}), \qquad PLS(\boldsymbol{X}, \boldsymbol{Y}) \equiv \{OLS(\boldsymbol{X}, \boldsymbol{y}^j)\}_{j=1,\ldots,q}. \tag{9}$$

La borne supérieure de la dimension du modèle, M, est donnée par le rang de $\boldsymbol{X}$, information utile lors du choix effectif de M par la validation croisée.
Signalons enfin une propriété qui fait le lien entre PLS et l'Analyse Exploratoire des Données, l'Analyse en Composantes Principales de X est l'auto-régression PLS de $\boldsymbol{X}$ sur lui même,

$$PLS(\boldsymbol{X}, \boldsymbol{Y} = \boldsymbol{X}) \equiv ACP(\boldsymbol{X}). \tag{10}$$

Le modèle PLS linéaire en les variables explicatives

D'après (5), une composante $\boldsymbol{t}^m$ est une combinaison linéaire des variables explicatives naturelles, les poids étant les éléments du vecteur $\boldsymbol{\theta}^m$. L'ensemble $\{\boldsymbol{\theta}^m\}_m$ est une famille ($\boldsymbol{V} = \boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}$)-orthogonale de l'espace vectoriel $Im\,\boldsymbol{X}'$

$$< \boldsymbol{t}^{m_1}, \boldsymbol{t}^{m_2} >_{\boldsymbol{D}} = < \boldsymbol{w}*^{m_1}, \boldsymbol{w}*^{m_2} >_{\boldsymbol{V}} = 0, \quad m_1 \neq m_2.$$

Le rôle de $\{\boldsymbol{\theta}^m\}_m$ est double :

- Projeter $\mathbb{V}$-orthogonalement les colonnes de $\boldsymbol{X}'$ sur $\{\boldsymbol{\theta}^m\}_m$ pour des visualisations 2-D du nuage des observations. Ces cartes des individus, très proches de la façon classique de "voir" les données dans les plans $(\boldsymbol{t}^i, \boldsymbol{t}^j)$, ont l'avantage de produire une mesure géométrique exacte de la qualité de la représentation d'une observation. Voir (Durand, 2002) pour une étude détaillée de ces photos bi-dimensionnelles.
- Exprimer récursivement les coefficients du modèle linéaire par rapport aux variables explicatives

  $$\hat{\boldsymbol{Y}}(M) = (\boldsymbol{H}_1 + \ldots + \boldsymbol{H}_M)\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}(M), \tag{11}$$

  où $\boldsymbol{\beta}(M)$ est la matrice $p \times q$ des coefficients du modèle,

$$\begin{aligned} \boldsymbol{\beta}(0) &= \boldsymbol{0} \\ \boldsymbol{\beta}(m) &= \boldsymbol{\beta}(m-1) + \boldsymbol{\theta}^m \boldsymbol{\theta}^{m\prime} \boldsymbol{X}'\boldsymbol{D}\boldsymbol{Y} / \|\boldsymbol{\theta}^m\|^2_{\mathbb{V}}, \qquad m = 1, \ldots, M. \end{aligned}$$

Le succès de PLS tient dans l'usage d'un petit nombre, M, de variables latentes non corrélées et facilement interprétables par les variables explicatives naturelles. Cette économie dans le nombre M de variables réellement explicatives rend le modèle linéaire robuste face aux problèmes de la rareté des observations et de la colinéarité entre certaines variables explicatives. En l'absence d'un jeu de données supplémentaires pour tester le modèle, le choix de M se fait par validation croisée sur l'échantillon d'apprentissage. Par exemple, une observation est enlevée pour être prédite par les données restantes. Ce processus est réitéré jusqu'à ce que toutes les observations aient été enlevées-prédites. La moyenne des carrés des erreurs de prédiction sur chaque réponse, le $PRESS$, PRredictive Error Sum of Squares, est le critère retenu pour mesurer la qualité du modèle. La séquence des valeurs de

$$PRESS(m) = \sum_{j=1}^{q} \frac{1}{n} \sum_{i=1}^{n} (y_i^j - < \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}(m)^j_{(-i)} >)^2, \qquad m = 1, \ldots, rang(\boldsymbol{X}),$$

permet de choisir la meilleure dimension M du modèle, en général, celle conférant au $PRESS$ la valeur minimum.

Le coût de calcul du $PRESS$ est parfois élevé, un critère de substitution est le GCV, ou validation croisée généralisée, qui est une pénalisation de la somme des carrés des résidus dont le calcul est instantané

$$GCV(\alpha, m) = \frac{\sum_{j=1}^{q} \frac{1}{n} \sum_{i=1}^{n} (y_i^j - < \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}(m)^j >)^2}{[1 - \alpha \frac{m}{n}]^2}.$$

Le produit αm est la mesure du nombre effectif de paramètres, i.e, la dimension réelle du modèle. Par défaut, le coefficient de la pénalité $\alpha = 1$ car $trace(\boldsymbol{H}_1 + \ldots + \boldsymbol{H}_m) = m$ est la dimension du modèle PLS linéaire.

4 PLS boostée par des bases de *B*-splines

Si la structure de PLS est celle d'un algorithme de boosting, la base d'apprentissage étant d'après (3) et (5) une combinaison linéaire des p variables explicatives $\boldsymbol{x} = (x^1, \ldots, x^p)$, l'objectif des méthodes de boosting est cependant la capture de non-linéarités et d'éventuelles interactions. L'idée la plus naturelle consiste à utiliser la propriété de linéarité de la base d'apprentissage pour l'enrichir par une extension de la dimension due à l'usage des bases de splines.

4.1 Quelques mots sur les splines de régression

À chaque variable x^i on associe un ensemble de fonctions $\mathcal{B}_i = \{B_1^i(.), \ldots, B_{r_i}^i(.)\}$ qui est une famille de B-splines (De Boor, 1978). L'autre famille de fonctions de base, les puissances tronquées, est envisageable mais les B-splines sont particulièrement bien adaptées au codage. Une spline est une combinaison linéaire

$$s^i(.) = \sum_{k=1}^{r_i} \theta_k^i B_k^i(.),$$

dont les coefficients sont à estimer par une méthode de régression. Les B-splines forment une base de l'espace vectoriel des fonctions splines sur l'intervalle $[x_-^i, x_+^i]$ de l'étendue des valeurs de x^i. Un élément de cet espace ou spline de régression, $s^i(.)$, est une séquence de r_i morceaux de polynômes, de même degré d_i, se raccordant en des points intérieurs à $[x_-^i, x_+^i]$ appelés les nœuds. Le raccordement a lieu avec un certain degré de régularité sur les dérivées contrôlé par la multiplicité du nœud concerné (De Boor, 1978). Si K_i est le nombre de nœuds, la dimension de l'espace spline $r_i = d_i + 1 + K_i$. Noter que $K_i = 0$ produit l'espace des polynômes de degré d_i sur l'intervalle des données.

Une B-spline est une fonction à support local, plus précisément, la B-spline $B_k^i(.)$ est nulle en dehors de son support noté S_k^i qui est un intervalle défini par deux nœuds distincts particuliers, voir (De Boor, 1978) (Durand, 2001) pour plus de détails. Le codage flou défini par les B-splines est l'application de $\mathbb{R}$ dans $\mathbb{R}^{r_i}$

$$x \rightarrow (B_1^i(x), \ldots, B_{r_i}^i(x)),$$

la valeur $B_k^i(x)$ s'interprète comme un indicateur de l'appartenance de x à S_k^i,

$$0 \leq B_k^i(x) \leq 1, \qquad \sum_{k=1}^{r_i} B_k^i(x) = 1.$$

Un choix judicieux des nœuds permet d'isoler l'influence sur $s^i(.)$ d'une valeur extrême de la variable qui ne perturbera la spline que par les $d+1$ B-splines dont le support contient cette valeur. Les B-splines confèrent aux modèles qui les utilisent un caractère de robustesse par rapport aux valeurs extrêmes des variables explicatives. La contrepartie de cet avantage est que $s^i(x) = 0,\ \forall x \notin [x_-^i, x_+^i]$ ce qui rend la prédiction efficace sur l'étendue de l'échantillon d'apprentissage.

4.2 Le modèle purement additif, PLSS

L'extension Partial Least-Squares Splines (Durand, 2001), en bref PLSS, consiste à utiliser dans PLS la base d'apprentissage

$$t = h(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{p} \sum_{k=1}^{r_i} \theta_k^i B_k^i(x^i) = \sum_{i=1}^{p} h_i(x^i)\,. \tag{12}$$

La dimension de la base d'apprentissage passe de p pour le modèle linéaire à $r = \sum_1^p r_i$ pour le modèle spline additif. Chaque composante PLSS s'exprime additivement par une somme de fonctions splines coordonnées et les graphiques $\{(x^i, h_i(x^i))\}_{i=1,\ldots,p}$, sont utilisés comme dans les Figures 7 et 8, pour interpréter l'influence des variables explicatives sur la variable latente t.

L'élaboration pas-à-pas du modèle est basée sur la matrice des bases de splines $\boldsymbol{B} = [\boldsymbol{B}_1|\ldots|\boldsymbol{B}_p]$, centrée en colonnes au sens de $\boldsymbol{D}$, le bloc $\boldsymbol{B}_i$ étant la matrice $n \times r_i$ dont les colonnes sont obtenues par codage de l'échantillon de x^i par la base de B-splines utilisée. Pour faire court,

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

Dans (11), le regroupement des termes par blocs d'indices d'une même famille de B-splines, fournit l'ajustement de la $j$ème réponse

$$\hat{\boldsymbol{Y}}^j(M) = [\boldsymbol{B}_1|\ldots|\boldsymbol{B}_p]\,\boldsymbol{\beta}^j(M) = \sum_{i=1}^{p} \boldsymbol{B}_i \boldsymbol{\beta}_i^j(M).$$

Les coefficients β des fonctions de base n'ont pas en général d'interprétation individuelle. L'exception est le cas d'une spline de degré 0 dont les fonctions de base sont constantes par morceaux à valeurs 0 ou 1 et conduisent à une matrice $\boldsymbol{B}_i$ de codage disjonctif complet. Alors, chaque coefficient β est le poids affecté au support concerné dont le sens est celui d'un niveau de valeur de la variable, par exemple, faible, moyen ou fort dans le cas de deux nœuds.

L'ajustement est celui d'un modèle multi-réponses de type purement additif (Hastie et Tibshirani, 1990),

$$\widehat{y}^j(M) = s_M^{j,1}(x^1) + \ldots + s_M^{j,p}(x^p), \qquad j = 1, \ldots, q. \tag{13}$$

L'influence non-linéaire des variables explicatives sur la $j$ème réponse est appréciée comme pour les composantes, par l'examen des graphes des courbes coordonnées $\{(x^i, s_M^{j,i}(x^i))\}_{i=1,\ldots,p}$

les plus significatives. Les variables explicatives étant standardisées, l'étendue de l'intervalle des valeurs de $s_M^{j,i}(\boldsymbol{x}^i)$ permet de classer les courbes coordonnées par ordre décroissant d'influence. Les propriétés (9) et (10) deviennent dans le cadre de PLSS :

$$\text{Si} \quad M = rang(\boldsymbol{B}), \qquad PLSS(\boldsymbol{X}, \boldsymbol{Y}) \equiv \{LSS(\boldsymbol{X}, \boldsymbol{y}^j)\}_{j=1,\ldots,q}, \tag{14}$$

i.e., PLSS produit les q régressions Least Squares Splines (Stone, 1985) lorsque celles-ci ont un sens (problème de l'inversibilité de la matrice $\boldsymbol{B}'\boldsymbol{D}\boldsymbol{B}$).

$$PLSS(\boldsymbol{X}, \boldsymbol{Y} = \boldsymbol{B}) \equiv PLS(\boldsymbol{B}, \boldsymbol{Y} = \boldsymbol{B}) \equiv ACP(\boldsymbol{B}), \tag{15}$$

i.e., l'ACP non-linéaire au sens de (Gifi, 1990) est l'auto-régression PLS de $\boldsymbol{B}$ sur lui même. Rappelons les paramètres de la méthode :

- La dimension du modèle M obtenue soit par validation croisée (PRESS) soit en utilisant la validation croisée généralisée (GCV).
- Pour chaque variable explicative, le degré, le nombre et l'emplacement des nœuds utilisés, i.e., les paramètres splines sur lesquels nous reviendrons.

4.3 Capture des interactions bi-variées par MAPLSS

Bien que PLS soit une méthode robuste face à la malédiction de la dimension, la capture d'interactions nécessite, à cause de l'explosion exponentielle des possibles, une phase de sélection des interactions significatives. C'est l'objectif des Multivariate Additive PLS Splines, en bref MAPLSS, (Durand et Lombardo, 2003) et (Lombardo et al., soumis), qui proposent une base d'apprentissage PLS introduisant des produit tensoriels de familles de B-splines prises deux à deux. Notons

$$\mathcal{I} = \{\{i, i'\} | \text{ l'interaction entre } x^i \text{ et } x^{i'} \text{ est acceptée}\}.$$

Le cardinal de $\mathcal{I}$ est un nombre compris entre 0, pas d'interaction, et $p(p-1)/2$, toutes les interactions bi-variées possibles sont acceptées. La base d'apprentissage MAPLSS devient

$$t = h(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{p} \sum_{k=1}^{r_i} \theta_k^i B_k^i(x^i) + \sum_{\{i,i'\} \in \mathcal{I}} \left[\sum_{k=1}^{r_i} \sum_{l=1}^{r_{i'}} \theta_{k,l}^{i,i'} B_k^i(x^i) B_l^{i'}(x^{i'}) \right]. \tag{16}$$

Le prix à payer pour la capture des interactions est l'extension de la dimension r de la base d'apprentissage qui devient

$$r = \sum_{i=1}^{p} r_i + \sum_{\{i,i'\} \in \mathcal{I}} r_i r_{i'}.$$

On comprend la nécessité de la sélection des interactions sur le "petit" exemple suivant : Vingt variables explicatives ($p = 20$) transformées par des B-splines de même dimension $r_i = 10$ et l'on désire incorporer toutes les interactions ($card(\mathcal{I}) = 190$), la dimension de la base d'apprentissage est $r = 19200$...

Dans MAPLSS, une composante t est une décomposition de type ANOVA des effets principaux et des interactions d'ordre 2 capturés par des fonctions splines uni-variées et bi-variées

$$t = \sum_{i=1}^{p} h_i(x^i) + \sum_{\{i,i'\} \in \mathcal{I}} h_{i,i'}(x^i, x^{i'}).$$

La matrice $\boldsymbol{B}$, $n \times r$ centrée en colonnes au sens de $\boldsymbol{D}$, est la matrice des nouvelles variables explicatives du codage par les bases de splines. Elle s'écrit

$$\boldsymbol{B} = [\boldsymbol{B}_1 | \ldots | \boldsymbol{B}_p \| \ldots | \boldsymbol{B}_{i,i'} | \ldots]$$

où $\boldsymbol{B}_{i,i'}$ est la matrice, $n \times r_i r_{i'}$, dont les colonnes sont les produits deux à deux des termes des colonnes de $\boldsymbol{B}_i$ et de $\boldsymbol{B}_{i'}$. En bref,

$$MAPLSS(\boldsymbol{X}, \boldsymbol{Y}) \equiv PLS(\boldsymbol{B}, \boldsymbol{Y}).$$

Le remplacement de $\boldsymbol{X}$ par $\boldsymbol{B}$ dans (11) conduit par le calcul des β, à l'ajustement des réponses selon la décomposition de type ANOVA, effets principaux plus interactions,

$$j = 1, \ldots, q, \qquad \hat{y}_M^j = \sum_{i=1}^{p} s_M^{j,i}(x^i) + \sum_{\{i,i'\} \in \mathcal{I}} s_M^{j,ii'}(x^i, x^{i'}). \tag{17}$$

De la même que dans PLSS, un coefficient β d'une base de splines n'est pas directement interprétable, une mesure de l'influence d'un terme de la décomposition est l'étendue des valeurs transformées par les fonctions splines. Selon ce critère il est instructif de classer par ordre décroissant d'influence les graphes 2-D et 3-D des courbes ou des surfaces de la décomposition (17). Grâce au classement précédent, l'obtention d'un modèle plus économique parfois de meilleure qualité de prédiction, est aisé à obtenir en élaguant (17) des termes les moins significatifs.

La construction du modèle

Entrées : $seuil = 20\%$, $M_{max} =$ nombre de dimensions à explorer.

0. Phase préliminaire : Le modèle PLSS des effets principaux repéré par m.

Choix des paramètres splines et de la dimension M_m. Noter $GCV_m(M_m)$ et $R_m^2(M_m)$ le CGV et le coefficient R^2.

1. Évaluation individuelle de toutes les $p(p-1)/2$ interactions.

Chaque interaction notée i est séparément ajoutée aux effets principaux.

$$CRIT(M_i) = \max_{M \in \{1, M_{max}\}} \frac{R_{m+i}^2(M) - R_m^2(M_m)}{R_m^2(M_m)} + \frac{GCV_m(M_m) - GCV_{m+i}(M)}{GCV_m(M_m)}$$

Éliminer les interactions telles que $CRIT(M_i) < 0$ et classer par ordre décroissant les interactions restantes candidates à entrer dans $\mathcal{I}$.

2. Ajouter successivement les interactions aux effets principaux (forward phase).

Soit $GCV_0 = GCV_m(M_m)$ et $i = 0$.
REPEAT
- $i \leftarrow i + 1$
- ajouter l'interaction i et indexer le nouveau modèle par i
- $GCV_i = \min_{M \in \{1, M_{max}\}} GCV_i(M)$

UNTIL $(GCV_i < GCV_{i-1} - seuil * GCV_{i-1})$

3. Élagage des termes ANOVA de plus faible influence (backward phase).

Critères : L'étendue des valeurs des fonctions ANOVA et le PRESS pour une mesure de la qualité de la prédiction et le choix de la dimension M.

La procédure précédente mise en œuvre dans MAPLSS vise à construire l'ensemble $\mathcal{I}$ des interactions significatives. Les paramètres d'entrée sont d'abord le *seuil* d'acceptation-rejet d'une interaction. Il est utilisé dans la phase 2 lors du calcul du gain relatif dans le GCV du modèle avec ou sans l'interaction concernée. Le seuil par défaut est 0.2, cette valeur est le résultat d'une campagne de simulations (Lombardo et al., soumis) basées sur trois types de signaux additifs, un signal purement additif, un signal avec une interaction et enfin un avec deux interactions. Le deuxième paramètre est le nombre maximum de composantes à explorer, il dépend des données mais doit être plus petit ou égal au rang des matrices de design.
La phase 0 préliminaire est obligatoire, elle consiste à construire le modèle PLSS des effets principaux. Ensuite, la phase 1 évalue séparément chaque interaction lorsqu'elle est ajoutée au effets principaux. Le critère, $CRIT$, est la somme de deux termes qui sont respectivement les gains relatifs dans le R^2 et dans le GCV. Lorsque PLSS produit un sur-ajustement aux données d'apprentissage, $R^2(M_m)$ est voisin de un et le premier terme de la somme est négligeable. À la fin de la phase 1, les interactions ayant un CRIT négatif sont éliminées, les autres sont classées par ordre décroissant du critère et forment les candidats à l'entrée dans l'ensemble $\mathcal{I}$ des interactions acceptées.

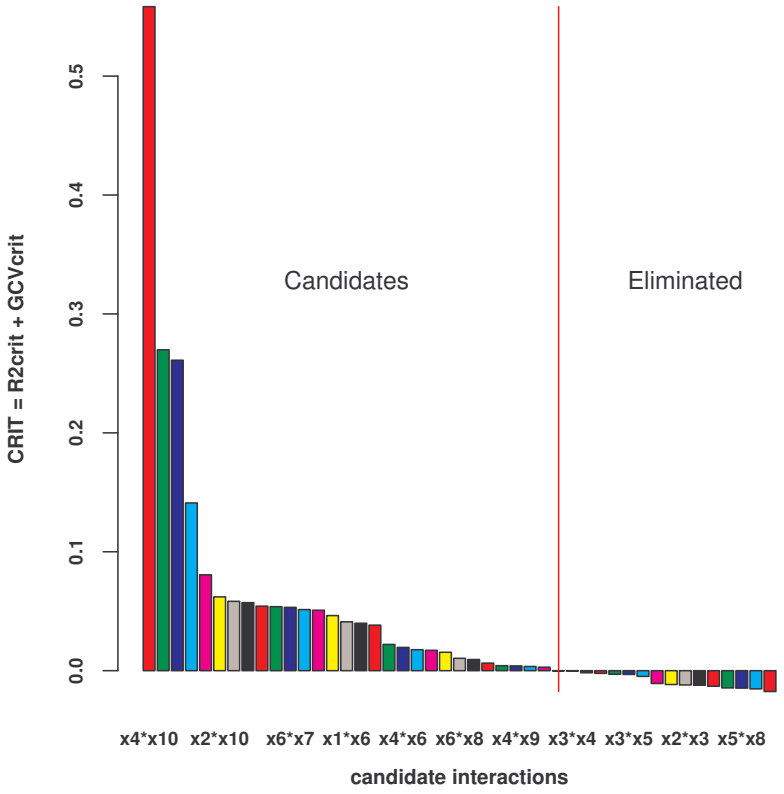


Figure 1 : Classement des interactions candidates à l'entrée dans le modèle de la réponse y^4 des données de polymérisation. L'interaction $x^4 * x^{10}$ arrive en tête. C'est la seule acceptée au seuil de 20% avec un gain du GCV de 52.18%.

La Figure 1 présente le classement des interactions candidates pour la modélisation de la réponse y^4 des données de polymérisation de la Section 6.2. L'interaction numéro 1 est $x^4 * x^{10}$, ce sera la seule à être définitivement acceptée dans la phase 2 avec un gain relatif du GCV de 52.18% (calculé avec 11 composantes et une pénalité $\alpha = 1.2$).
Après avoir ordonné les termes de la décomposition (17) par ordre décroissant selon l'amplitude de l'intervalle des valeurs de chaque fonction ANOVA, la phase 3, ultime, de la procédure consiste à élaguer les termes négligeables. La validation croisée et le PRESS sont utilisés pour valider ces modèles plus économiques.

5 Stratégies de choix des splines de régression

La capture des non-linéarités et des interactions nécessite un soin particulier dans l'utilisation des splines de régression. D'une part, le choix optimal des nœuds est un problème difficile même dans le cas de la régression simple, citons, entre autres tentatives (Molinari et al., 2004). On peut cependant penser que l'optimalité n'est pas indispensable pour un résultat satisfaisant, dans le cas uni-varié comme dans le cas multi-varié. Pour combattre le sur-ajustement aux données en régression simple, l'idée efficace de Eilers et Marx (1996) consiste à choisir plus de nœuds (équidistants) que nécessaire et pénaliser la fonction objectif des moindres carrés par des différences finies portant sur les coefficients des B-splines adjacentes. Le degré étant choisi a priori, les P-splines ainsi définies, dépendent d'un seul paramètre de pénalisation.

Le problème se complique dans le cadre multi-varié et l'approche la plus connue est celle de MARS, (Friedman, 1991), qui propose un choix adaptatif des splines de base, les puissances tronquées, ainsi que des nœuds. Dans le logiciel *PLSS*, aucune procédure automatique de calcul des paramètres splines n'est utilisée pour les fonctions PLSS et MAPLSS ce qui est un inconvénient et un avantage. Un inconvénient, car le choix manuel du degré et des nœuds peut être laborieux pour un nombre élevé de variables explicatives. Dans ce cas, comme première approche, le choix des nœuds équidistants ou des nœuds aux quantiles est une option raisonnable. Un avantage, car sélectionner judicieusement degrés et nœuds permet à l'utilisateur de mettre en application ses connaissances des phénomènes ayant contribué à l'élaboration des données, (Avelino et al., 2007).

La détermination des paramètres splines s'inscrit dans une démarche de type Data Mining et consiste en un aller-retour entre le choix du type de modèle et sa validation. Deux stratégies, (Durand, 2001), peuvent être utilisées pour trouver un équilibre entre *économie* de la dimension (M du modèle et r de la base d'apprentissage) et *qualité* de la prédiction ($PRESS$ ou GCV).

La stratégie ascendante, consiste à augmenter progressivement le degré et les nœuds. Par défaut, le degré est égal à un et aucun nœud n'est utilisé, ce qui produit un modèle PLSS linéaire, point de départ raisonnable dans la recherche d'un modèle additif idéal. On peut explorer les modèles polynômiaux en augmentant les degrés tout en ne prenant aucun nœud. Enfin, plus de souplesse s'obtient par l'introduction de nœuds sachant qu'ajouter un nœud augmente la flexibilité locale d'une spline et ainsi, la liberté d'ajuster les données dans cette région.

Au contraire, la stratégie descendante commence par un degré élevé, trois par exemple, et plus de nœuds que "nécessaire". Ensuite, les nœuds superflus sont enlevés et le degré est diminué autant que possible.

Au cours de la campagne d'essais, l'évolution des graphes des fonctions ANOVA significatives (critère de l'étendue des valeurs de la fonction) et celle du $PRESS(M)$ optimal sont les indicateurs pour l'arrêt de la stratégie utilisée.

6 Exemples d'application

Trois jeux de données vont servir à illustrer la pratique et les performances de la régression PLS boostée. Les deux premiers présentent des variables dépendantes continues qui sont modélisées par des régressions uni-réponses, le troisième utilise la régression multi-réponses pour diagnostiquer l'appartenance à un groupe dans le cadre d'une classification avec superviseur. Tous les résultats numériques et graphiques ont été obtenus par le logiciel *PLSS*.

6.1 Un jeu de données simulées

Le jeu de données est extrait d'une campagne de simulations (Lombardo et al., soumis) dont les résultats sont rapportés dans la conférence du cycle J.P. Fénelon. L'objectif a été de comparer les performances des trois méthodes, BRUTO (Hastie et Tibshirani, 1990), MARS (Friedman, 1991) et MAPLSS, toutes basées sur les splines et produisant des modèles additifs. Neuf cents jeux de données ont été construits de la façon suivante. Trois types de signaux, l'un purement additif, le second incorporant une interaction, le troisième deux interactions. Trois nombres d'observations différents, cinquante, cent puis deux cents. Enfin, cent jeux de données aléatoires pour chacun des neuf cas précédents ont permis de tester les trois méthodes. Il ressort de ces comparaisons que BRUTO est la meilleure dans le cas d'un signal purement additif et un nombre d'observations assez grand ($n \geq 100$)), et que MAPLSS l'emporte en présence d'interactions sur des échantillons de taille moyenne et petite ($n = 100,\ n = 50$).

L'exemple utilise la fonction signal numéro deux de la série d'essais précédente associée à cent observations. L'objectif est de détailler les étapes de MAPLSS et la présentation des résultats. Le jeu de données d'apprentissage $(\boldsymbol{x}_i, y_i)_{i=1,\ldots,100}$ dérive du modèle $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$, $\boldsymbol{x}_i \sim U[0,1]^{10}$ et $\varepsilon_i \sim N[0,1]$, et du signal

$$f(\boldsymbol{x}) = 10\sin(\pi x^1 x^2) + 20(x^3 - 0.5)^2 + 10x^4 + 5x^5 + 0\ \textstyle\sum_{j=6}^{10} x^j.$$

La précision du modèle MAPLSS de dimension M est mesurée sur un autre jeu de données de mêmes dimensions, $(\boldsymbol{x}_i, y_i = f(\boldsymbol{x}_i))_{i=1,\ldots,100}$ grâce à l'erreur quadratique moyenne $MSE(M) = 100^{-1}\sum_{i=1}^{100}(f(\boldsymbol{x}_i) - \hat{y}_i(M))^2$. Les splines retenues sont présentées Table 2.

Table 2 : Degrés et nœuds des splines pour le jeu des données simulées.

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8	x^9	x^{10}
degré	1	1	2	1	1	1	1	1	1	1
nœuds	0.5	0.5	0.5							

MAPLSS détecte une interaction significative, $x^1 * x^2$, avec un gain relatif de 56.72% sur le GCV utilisant la pénalité $\alpha = 1$ et $M = 6$ composantes.
Suivant le critère de l'étendue des valeurs des splines, le classement des effets par ordre décroissant est le suivant :

$\boldsymbol{x}^4$	$\boldsymbol{x}^1 * \boldsymbol{x}^2$	$\boldsymbol{x}^3$	$\boldsymbol{x}^5$	$\boldsymbol{x}^1$	$\boldsymbol{x}^2$	$\boldsymbol{x}^{10}$	$\boldsymbol{x}^7$	$\boldsymbol{x}^9$	$\boldsymbol{x}^8$	$\boldsymbol{x}^6$
1.75962	1.31994	0.95828	0.95438	0.79419	0.76246	0.14731	0.08212	0.0694	0.06561	0.05034

Il est clair que l'on peut éliminer les cinq derniers du classement qui sont justement les effets des cinq variables qui n'apportent que du bruit dans le modèle. Ici, deux critères de la qualité de la prédiction sont utilisables et permettent le choix de la dimension M, l'un interne aux données d'apprentissage, le $PRESS$, l'autre externe construit sur l'échantillon test, le MSE. Les cinq derniers termes enlevés, les valeurs optimales obtenues sont $PRESS(0.01, 6) = 0.0622$ (1% des données enlevées-prédites, $M = 6$) et $MSE(8) = 0.382$. La question maintenant posée est la suivante, doit-on continuer l'élagage par le bas des deux candidats suivants qui sont les effets principaux x^2 et x^1 ? Ces effets ne sont pas dans le signal qui dépend seulement de l'interaction $x^1 * x^2$. Enlevons les effets principaux x^1 et x^2, les valeurs optimales des critères sont : $PRESS(0.01, 3) = 0.072$ et $MSE(8) = 0.383$. Le n'est pas significatif, il est même légèrement négatif.

En résumé, MAPLSS a capturé l'interaction $x^1 * x^2$, éliminé le bruit des cinq dernières variables, mais n'a pas retrouvé exactement les termes du signal puisque les effets x^1 et x^2 entrent dans le modèle. La qualité de la prédiction est excellente comme le montre la Figure 2 et la valeur de l'erreur relative moyenne optimale obtenue avec 8 dimensions, $MRE(8) = \frac{1}{100}\sum_{i=1}^{100}\left|\frac{f(\boldsymbol{x}_i)-\hat{y}_i(8)}{f(\boldsymbol{x}_i)}\right| = 0.0345$.

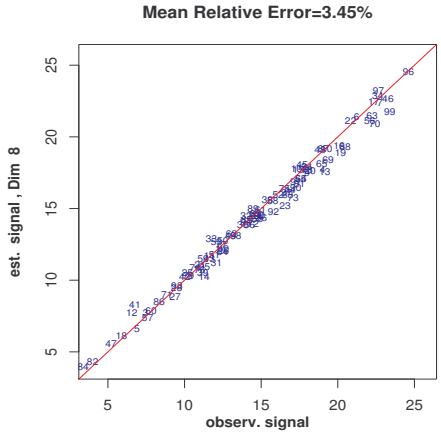


Figure 2 : Prédiction MAPLSS à 8 dimensions sur l'échantillon test des données simulées : valeurs prédites contre valeurs du signal.

6.2 Les données de polymérisation

Le jeu de données "chemdata" (De Veaux et al., 1993) comprends $p = 10$ variables explicatives et quatre réponses à valeurs dans $[0, 1]$. Le nombre d'observations est $n = 61$. Il s'agit de tests de polymérisation dans une expérience pilote dont la confidentialité ne permet pas de donner plus de détails. Ces données ont été utilisées dans (De Veaux et al., 1993) pour comparer les méthodes neuronales et MARS dans la prédiction séparée des quatre réponses. La modélisation par MAPLSS de la première réponse, y^1, est dans (Lombardo et al., soumis) et nous présentons ici la comparaison des résultats de MAPLSS et de MARS pour les quatre réponses. Seule la décomposition de type ANOVA du modèle MAPLSS de y^4 est commentée.

Le critère pour évaluer la qualité de la prédiction est le PRESS obtenu par validation croisée en enlevant une observation à la fois. La Figure 3 représente les dix nuages $(x^j, y^4)_{j=1,\ldots,10}$ des 61 observations, lissés par les splines aux moindres carrés avec les degrés et les nœuds présentés dans la Table 3.

Table 3 : Degrés et nœuds des splines pour les variables de "chemdata".

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8	x^9	x^{10}
degré	1	1	0	1	1	1	1	1	1	3
nœuds	0.6		0.2		0.3		0.3			

Ces même jeu de paramètres splines a été utilisé dans PLSS et MAPLSS pour la modélisation des quatre réponses qui sont deux par deux très fortement corrélées, (y^1, y^2) et (y^3, y^4). La Table 4 résume les valeurs du PRESS obtenues par PLS linéaire classique, par le modèle PLSS purement additif et par MAPLSS qui a introduit l'interaction $x^4 * x^{10}$ dans les modèles des quatre réponses. La prédiction par MAPLSS est excellente et le gain par rapport à PLS et PLSS est spectaculaire. Les résultats de MARS dans l'article (De Veaux et al., 1993) sont rapportés dans la dernière colonne de la Table 4. MARS a détecté et incorporé dans le modèle l'interaction $x^4 * x^{10}$ et à un degré moindre $x^6 * x^{10}$ ainsi que l'interaction $x^7 * x^{10}$.

Table 4 : Dimension du modèle et valeur du PRESS (1 out) pour la prédiction des 4 réponses de "chemdata" par PLS linéaire, PLSS et MAPLSS sans élagage des effets négligeables. Le PRESS de MARS est dans la dernière colonne.

	PLS	PLSS	MAPLSS	MARS
y^1	(2, 0.5709)	(3, 0.4835)	(7, 0.0584)	0.0106
y^2	(3, 0.6094)	(3, 0.5760)	(7, 0.2108)	0.1114
y^3	(5, 0.1595)	(5, 0.1478)	(14, 0.0626)	0.006
y^4	(4, 0.1268)	(7, 0.1267)	(14, 0.0671)	0.014

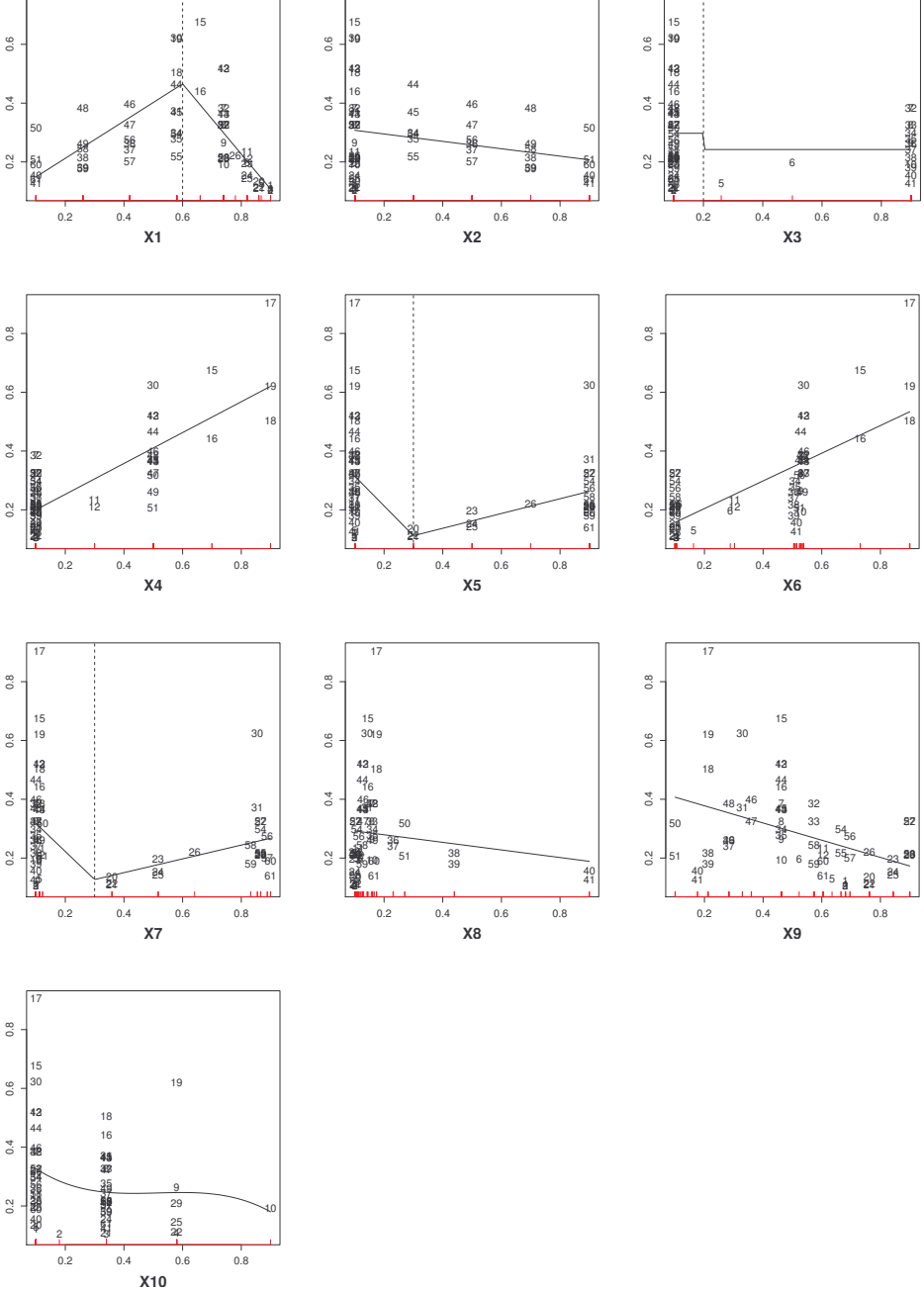


Figure 3 : Les 10 nuages de points $(x^j, y^4)_{j=1,10}$ du jeu de données "chemdata", lissés par des splines aux moindres carrés. Les mêmes degrés et nœuds, repérés ici par les verticales en pointillés, sont utilisés dans PLSS et MAPLSS pour la modélisation des quatre réponses $y^1, \ldots, y^4$.

C'est sûrement l'une des causes des meilleurs résultats de MARS qui sont cependant du même ordre que ceux obtenus par MAPLSS, sauf pour la réponse y^3, 10^{-3} pour MARS et 10^{-2} pour MAPLSS.

La Figure 4 représente la décomposition de type ANOVA pour la prédiction de y^4 par MAPLSS. Les effets principaux et les interactions sont classés par ordre décroissant d'importance, de gauche à droite et de haut en bas, avec pour critère l'étendue des valeurs transformées par les splines uni-variées ou bi-variées.

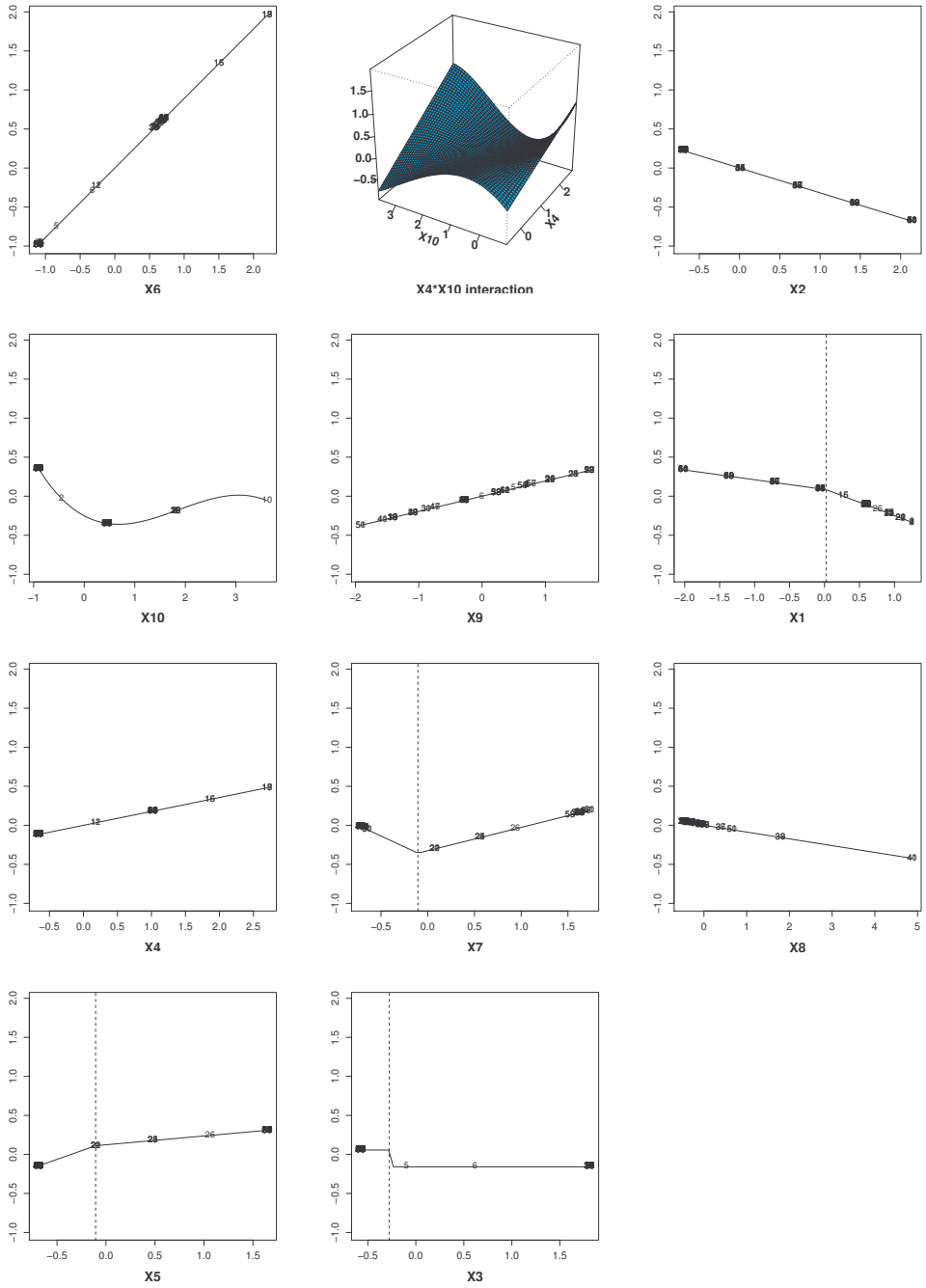
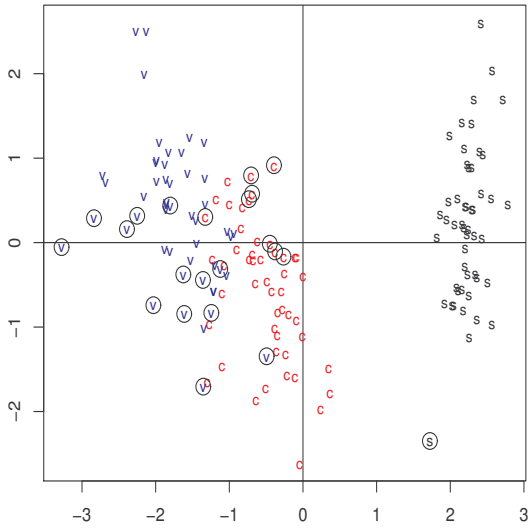


Figure 4 : La décomposition de type ANOVA de y^4. Classement décroissant, de gauche à droite et de haut en bas, selon l'étendue des valeurs sur l'axe vertical.

Notons que la phase d'élagage des effets de moindre influence n'a pas apporté une amélioration significative du PRESS et que le modèle MAPLSS de la Figure 4 contient tous les effets principaux plus l'interaction $x^4 * x^{10}$. Sur chaque figure, les valeurs, en abscisses, des variables explicatives sont centrées-réduites et leurs valeurs transformées par les splines, en ordonnées, sont centrées. Le modèle fait ressortir l'effet linéaire de x^6 et l'interaction $x^4 * x^{10}$ qui apporte à elle seule le gain de prédiction par rapport au modèle PLSS des effets principaux.

6.3 Les iris de Fisher revisités par PLS boostée

La longueur et la largeur des pétales, la longueur et la largeur des sépales ont été mesurées sur cinquante spécimens de chacune des trois espèces, Iris setosa (s), I. versicolor (c) et I. virginica (v). Les dimensions des données d'apprentissage sont ici, $n = 150$, $p = 4$ et $q = 3$ pour le nombre de colonnes de la matrice $\boldsymbol{Y}$ du codage disjonctif complet indiquant l'appartenance des individus à l'un des trois groupes. C'est le choix typique des réponses pour la régression PLS discriminante linéaire, (Sjöström et al., 1986), (Tenenhaus, 1998), voir aussi (Sabatier et al., 2003) pour l'usage de la métrique de Mahalanobis dans une généralisation de la régression PLS discriminante.



PLS linéaire

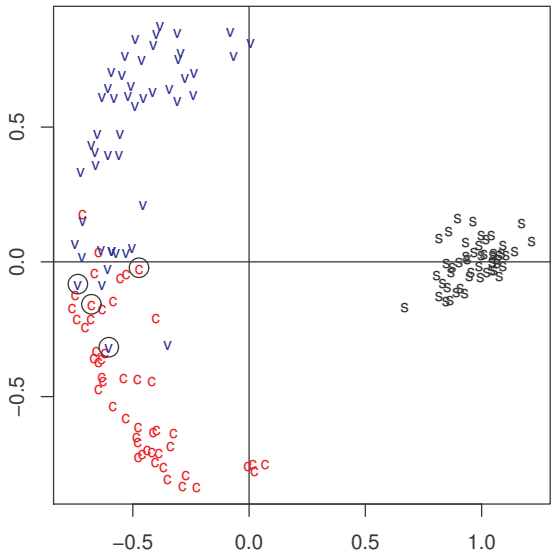
Table des diagnostics

	s réel	c réel	v réel
s prédit	**49**	0	0
c prédit	1	**42**	13
v prédit	0	8	**37**

Figure 5 : À gauche, le nuage des observations des Iris setosa (s), versicolor (c) et virginica (v), dans le plan (t^1, t^2) de PLS linéaire. Les individus mal affectés à leurs groupes réels sont entourés d'un cercle. À droite, les diagnostics d'affectation des observations, les individus biens classés sont en gras.

Les résultats de PLS linéaire classique, $PRESS(3) = 1.379$, (10% out), sont résumés Figure 5 qui montre d'une part le rôle des deux variables discriminantes t^1 et t^2 dans la discrimination des groupes et d'autre part la table des diagnostics d'affectation obtenue géométriquement en classant un individu de la matrice des composantes principales dans le groupe le plus voisin. Plus précisément, le critère est la distance de Mahalanobis entre un individu et l'individu moyen du groupe.

Le boosting PLS mis en oeuvre sur cet exemple fait ressortir le modèle PLSS purement additif, $PRESS(4) = 0.3551$, (10% out), basé sur des splines de degré deux avec trois nœuds équidistants. La Figure 6 montre le gain dans la qualité du diagnostic par rapport à PLS linéaire. La deuxième composante t^2 sépare mieux que son homologue PLS linéaire, les deux groupes versicolor et virginica et quatre individus sont mal classés. Les résultats sont proches de ceux obtenus sur cet exemple, avec les mêmes paramètres splines, par l'Analyse en Composantes Principales sur Variables Instrumentales, ou ACPVI, (Durand, 1993). Le contexte des données sur les iris, petit nombre de variables avec suffisamment d'observations, est bien adapté à l'ACPVI qui est moins robuste que PLS face à la malédiction de la dimension et qui produit ici trois erreurs de diagnostic.



PLS boostée

Table des diagnostics

	s réel	c réel	v réel
s prédit	**50**	0	0
c prédit	0	**48**	2
v prédit	0	2	**48**

Figure 6 : À gauche, le nuage des observations des Iris setosa (s), versicolor (c) et virginica (v), dans le plan (t^1, t^2) du modèle PLSS. Les individus mal affectés à leurs groupes réels sont entourés d'un cercle. À droite, les diagnostics d'affectation des observations, les individus biens classés sont en gras.

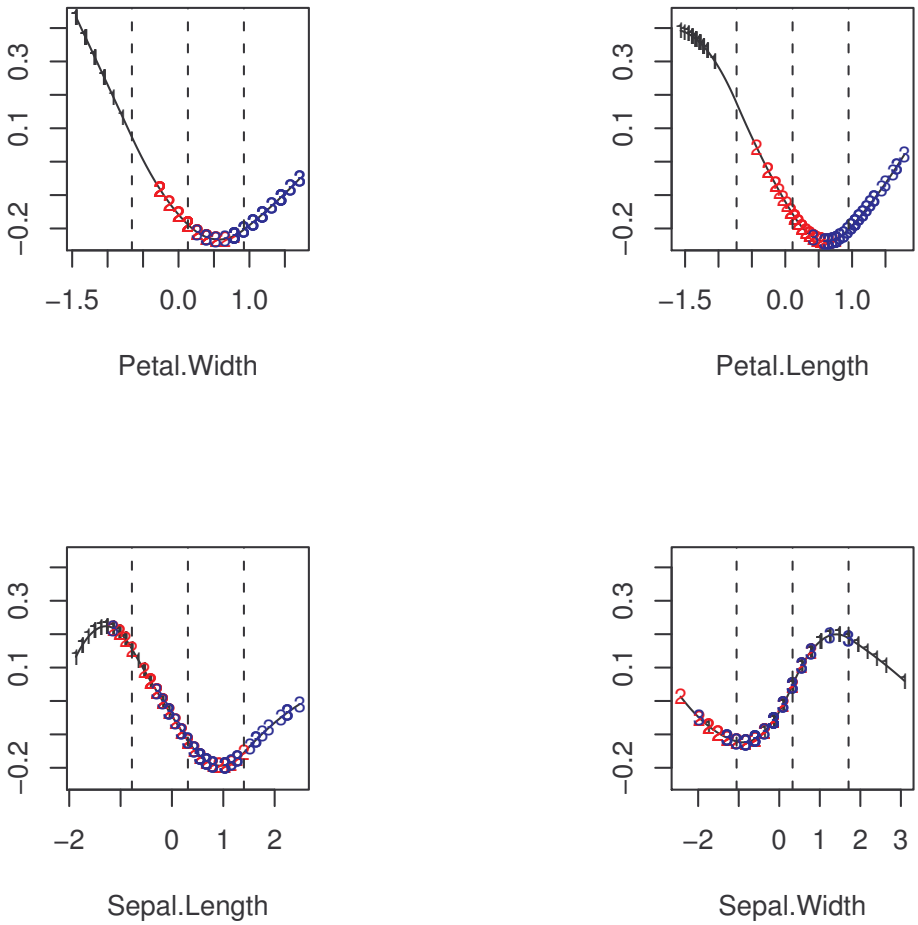


Figure 7 : La décomposition de type ANOVA pour t^1, les variables sont ordonnées par ordre décroissant d'influence, de la gauche vers la droite et du haut vers le bas, selon l'étendue des valeurs sur l'axe vertical. Le codage des observations : setosa (1), versicolor (2), virginica (3). Les verticales en pointillés indiquent la position des noeuds.

Si l'on s'intéresse au rôle des variables naturelles dans la discrimination il est utile d'examiner les graphiques des fonctions splines de la décompostion (12) d'une variable discriminante t. Rappelons que t^1 sépare le groupe setosa des deux autres et que t^2 discrimine versicolor et virginica. D'après la Figure 7 montrant la décomposition de t^1, setosa, groupe 1, se différencie surtout par de faibles dimensions des pétales et à un degré moindre de la longueur des sépales.

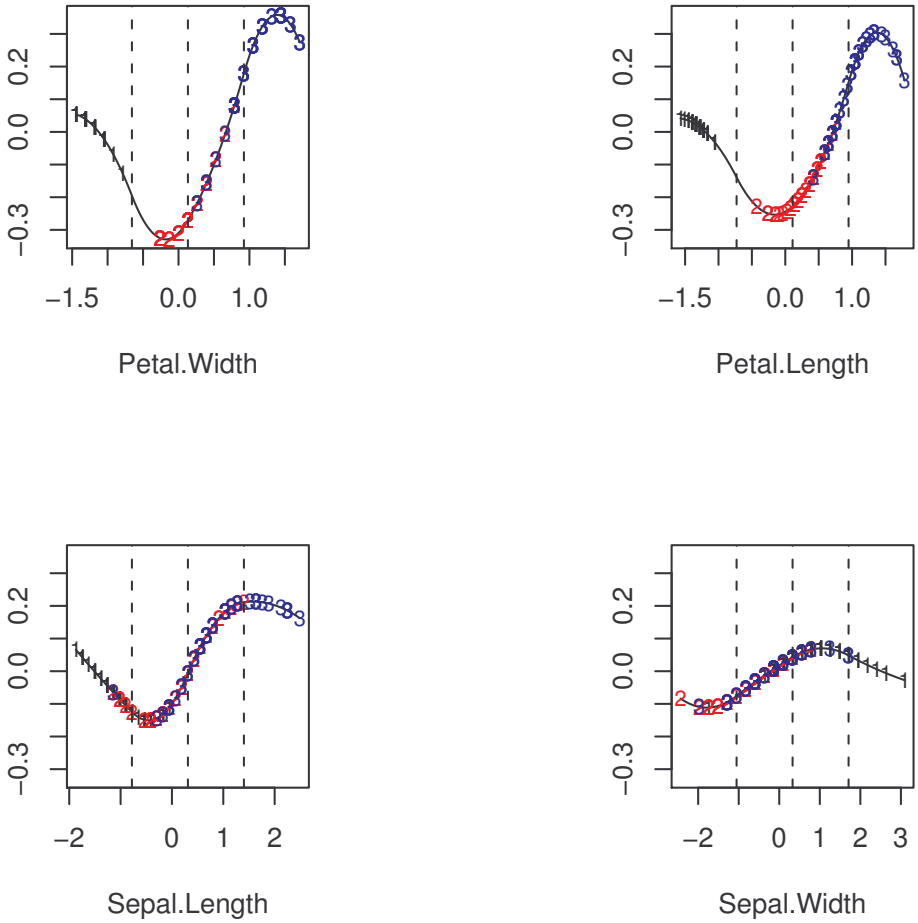


Figure 8 : La décomposition de type ANOVA pour t^2, les variables sont ordonnées par ordre décroissant d'influence, de la gauche vers la droite et du haut vers le bas, selon l'étendue des valeurs sur l'axe vertical. Le codage des observations : setosa (1), versicolor (2), virginica (3). Les verticales en pointillés indiquent la position des noeuds.

La décomposition de t^2, Figure 8, montre que le groupe 3, virginica, présente de fortes dimensions des pétales alors qu'elles sont seulement moyennes pour le goupe 2, versicolor, (le zéro sur les axes est la valeur moyenne). En résumé, les variables les plus discriminantes sont la largeur et la longueur des pétales et à un degré moindre la longueur des sépales.

7 Conclusion

La régression PLS linéaire est une méthode robuste face au problème de la rareté des observations et de la colinéarité des variables explicatives. Telle qu'elle est apparue dans les dernières décennies du XXième siècle, c'est une des premières méthodes de type boosting L_2 avant même que le boosting ne fasse l'objet des formalisations théoriques récentes et des applications pratiques reconnues comme efficaces. Cet article resitue PLS dans ce contexte en présentant une variable latente comme base d'apprentissage. Une des originalités du boosting PLS est qu'il utilise des pseudo variables explicatives, faisant ainsi la liaison entre les méthodes de régression et celles de l'Analyse Exploratoire des Données, l'Analyse en Composantes Principales étant l'auto-régression PLS de $\boldsymbol{X}$ sur lui même.

L'usage des bases de B-splines dans la construction de la base d'apprentissage exploite tout le potentiel du boosting PLS en permettant aux modèles additifs PLSS et MAPLSS de capturer

non-linéarités et interactions bi-variées. Le prix à payer est l'extension de la dimension de la base d'apprentissage bien supportée par PLS ce qui rend ces méthodes compétitives même dans le cas de petits échantillons d'apprentissage. L'introduction d'interactions d'ordre supérieur à deux dans la décomposition de type ANOVA du modèle MAPLSS ne présente pas de difficultés théoriques réelles. Cependant, deux raisons ont fait qu'il a semblé plus réaliste d'arrêter à deux l'ordre des interactions. La première est la lourdeur de l'exploration de l'arbre des possibles, la deuxième est le manque d'interprétation des interactions supérieures.

Enfin, les propriétés des B-splines en font un outil attractif. D'une part, le support local rend les modèles robustes face aux valeurs extrêmes des variables explicatives. D'autre part, les propriétés de codage flou des données, déjà exploitées en Analyse Exploratoire des Données, trouveront sûrement, grâce à PLS, des applications non encore exploitées.

Références

[1] Avelino J., Cabut S., Barboza B., Barquero M., Alfaro R., Esquivel C., Durand J.F., et Cilas C., *Topography and crop management are key factors for the development of american leaf spot epidemics on coffee in Costa Rica*. Phytopathology, 97, pages 1532-1542, 2007.

[2] Breiman L., *Prediction games and arcing algorithms*. Neural Computation, 11, pages 1493-1517, 1999.

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

[4] Bühlmann P. et Bin Yu, *Boosting with the L_2 Loss : Regression and Classification*. Journal of the American Association, 98, pages 324-339, 2000.

[5] De Veaux R.D., Psichoggios D.C. et Ungar L.H., *A comparison of two nonparametric estimation schemes : MARS and neural networks*. Computers chem. Engng, Vol. 17, No. 8, pages 819-837, 1993.

[6] Durand J.F., *Generalized principal component analysis with respect to instrumental variables via univariate spline transformations*. Computational Statistics & Data Analysis, 16, pages 423-440, 1993.

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

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

[9] Durand J.F. et Lombardo R., *Interactions terms in nonlinear PLS via additive spline transformations*. In M. Schader, W. Gaul, M. Vichi (Eds), Stu-

dies in Classification, Data Analysis, and Knowledge Organization, Between Data Science and Applied Data Analysis, Springer-Verlag, pages 22-29, 2003.

[10] Eilers P.H.C. et Marx B.D., *Flexible smoothing with B-splines and Penalties, (with discussion)*. Statistical Science, 19, pages 89-121, 1996.

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

[12] Friedman J.H., *Greedy function approximation : a gradient boosting machine*. The Annals of Statistics, 29, 5, pages 1189-1232, 2001.

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

[14] Hastie T. et Tibshirani R., *Generalized additive models*. Chapman and Hall, London, 1990.

[15] Hastie T., Tibshirani R., et Friedman J.H., *The Elements of Statistical Learning*, Springer, 2001.

[16] Lombardo R., Durand J.F. et De Veaux R.D., *Multivariate Additive Partial Least-Squares Splines, MAPLSS*. Soumis.

[17] Molinari N., Durand J.F. et Sabatier R., *Bounded Optimal knots for Regression Splines*. Computational Statistics & Data Analysis, 2, pages 159-178, 2004.

[18] R Development Core Team, *R : A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006.

[19] Sabatier R., Vivien M. et Amenta P., *Two approaches for discriminant Partial Least-Squares*. In M. Schader, W. Gaul, M. Vichi (Eds), Studies in Classification, Data Analysis, and Knowledge Organization, Between Data Science and Applied Data Analysis, Springer-Verlag, pages 100-108, 2003.

[20] Sjöström M., Wold S. et Söderström B., *PLS discrimination plots*. In E.S. Gelsema et L.N. Kanals (Eds) : Pattern Recognition in Practice II. Elsevier, Amsterdam, 1986.

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

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

[23] Tukey J.W., *Exploratory Data Analysis*. Reading Massachusetts : Addison-Wesley, 1977.

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