

Résumé : Une récente extension du modèle Partial Least Squares linéaire (PLS) au modèle spline additif (PLSS) est présentée sous forme d'un guide d'utilisation S-Plus de la version 9.6 du logiciel programmé par l'auteur. PLSS hérite des avantages du modèle PLS face aux problèmes de la rareté des observations et de la présence de fortes corrélations entre les variables explicatives. En outre, PLSS propose à l'utilisateur un large éventail de modèles PLS allant du modèle linéaire à la régression par des polynômes locaux en passant par la modélisation polynomiale traditionnelle. L'usage des bases de B -splines à support local, rend les modèles PLSS plus robustes face aux problèmes des valeurs extrêmes des variables explicatives. L'exemple, basé sur un jeu de données sensorielles, est typique des problèmes précédents. Il permet d'illustrer les propriétés de PLSS et d'interpréter les modèles additifs obtenus. L'accent est mis sur le choix et le rôle des super-paramètres de la méthode.

Abstract: This paper aims at presenting as a user's guide, S-Plus version 9.6 programmed by the author, the recently devised extension of the Partial Least Squares (PLS) linear regression to the nonlinear additive spline model (PLSS). This method inherits from standard PLS of the well known advantages of the model in the presence of few observations and strong correlation in the predictors. Moreover, PLSS provides the user with PLS modeling tools that range from the linear to the local polynomial models, including classical polynomial modeling. The use of B -splines basis functions with local support, makes PLSS models more robust in the presence of extreme values of the predictors. The illustration of the PLSS properties as well as an interpretation of the additive models are made on a data set from sensometrics which is typical of such problems. A discussion on the role of the PLSS tuning parameters is proposed all along this report.

Contents

1	Introduction	3
2	Les splines de régression sous S-Plus	5
2.1	La fonction bs sous S-Plus	6
2.1.1	Le graphe des fonctions de base	7
2.1.2	Propriétés des <i>B</i> -splines	8
2.1.3	Exercices sous S-Plus	10
2.2	Les splines aux moindres carrés en régression simple	12
2.2.1	Les données des jus d'orange	13
2.2.2	La régression LSS entre <i>Chargé</i> et <i>Conductivité</i>	14
2.2.3	Exercices sous S-Plus	18
2.3	Le modèle spline additif dans le cadre multivarié	19
3	La régression PLS Spline (PLSS)	22
3.1	Le modèle PLS (linéaire)	22
3.2	Le modèle PLSS	24
4	Pratique de PLSS	26
4.1	La régression de <i>Chargé</i>	27
4.1.1	L'option par défaut de PLSS : PLS linéaire	27
4.1.2	Augmentation du degré et du nombre de noeuds	34
4.2	Régression multi-réponses de l' <i>Intensité</i> et de l' <i>Agressivité</i>	38
4.2.1	Prédiction-validation sur un jeu de données test	38
4.2.2	Exploration des données et modélisation	40
5	Quelques dernières remarques	44
6	ANNEXE	45
6.1	Annexe 1 : Données des jus d'orange	45
6.2	Annexe 2 : Description des fonctions S-Plus implémentées	46
6.3	Annexe 3 : Un peu de Mathématique sur les B-splines	48
7	BIBLIOGRAPHIE	50

1 Introduction

Ce travail fait suite au guide d'utilisation (Durand, Roman et Vivien 1998) de la régression Partial Least Squares linéaire (PLS) développée dans le langage statistique S-plus© (MathSoft 1996) par J.F. Durand et R. Sabatier. La régression PLS due à H. Wold (1975) et à S. Wold, H. Martens et H. Wold (1983) pour la version multi-réponses, est maintenant une méthode standard d'exploration et de modélisation linéaire des données. On peut se référer à Tenenhaus (1998) et (Tenenhaus, Gauchi et Ménardo 1995) pour un exposé en français des propriétés et des perspectives de cette approche.

La première extension de PLS au modèle non linéaire additif, appelée Additive Spline PLS ou ASPLS (Durand, Sabatier 1997) est une méthode itérative qui utilise les splines ou polynômes par morceaux, comme dans l'extension non linéaire de l'Analyse en Composantes Principales sur Variables Instrumentales (Durand 1993).

La deuxième extension non linéaire de PLS faisant l'objet de ce rapport, est appelée PLS Spline ou PLSS (Durand 1997, 1999) et produit comme la première, un modèle spline additif. Elle présente l'avantage de mettre en oeuvre une méthode non itérative au sens où les composantes principales sont obtenues comme dans PLS par une technique de valeurs propres-vecteurs propres. En fait, PLSS n'est pas autre chose que la régression PLS linéaire entre la (ou les) réponse(s) Y et la matrice B du codage de l'échantillon X des variables explicatives ou prédicteurs par les fonctions de base de l'espace vectoriel des fonctions splines. Le prix à payer pour le passage au non linéaire étant l'expansion de la dimension colonne de la nouvelle matrice de "design" B

$$PLSS(X, Y) = PLS(B, Y).$$

La méthode PLSS supporte remarquablement bien l'expansion de la dimension qui peut être un inconvénient rédhibitoire pour d'autres approches comme celle des splines aux moindres carrés (LSS) sur certains jeux de données.

L'utilisateur averti de PLS linéaire connaît l'avantage principal de la méthode : validité du modèle obtenu, même dans les cas où peu d'observations sont disponibles et/ou lorsque certaines variables explicatives sont très corrélées. Cependant, comme pour d'autres méthodes factorielles de régression, les observations extrêmes sur certaines variables explicatives, ont une influence globale sur les composantes PLS et donc sur le modèle linéaire.

L’usage de la base des B -splines est un atout pour PLSS dans ce contexte. Ces fonctions à support local, nulles en dehors d’un certain intervalle, garantissent une influence locale des données extrêmes des prédicteurs sur le modèle. Le jeu de données qui sert d’application aux fonctions S-Plus programmées par l’auteur, est typique de ce problème.

Une autre particularité de l’espace vectoriel des fonctions splines fait de PLSS un modèle de régression capable de capturer aussi bien des relations linéaires, polynomiales que des relations polynomiales par morceaux. En effet, l’espace vectoriel des polynômes de degré fixé sur un intervalle $[a, b]$, est un espace spline particulier pour lequel aucun point de raccordement entre deux polynômes adjacents - un tel point est appelé un “noeud” - n’existe. L’option par défaut de la fonction `plss` utilise comme base celle des polynômes du premier degré sur $[a, b]$ ce qui produit un modèle PLS linéaire. Ainsi, l’utilisateur de PLSS dispose à son gré, d’une batterie d’outils de modélisation PLS allant du modèle linéaire au modèle polynomial par morceaux en passant par la modélisation polynomiale usuelle sur un intervalle $[a, b]$.

A travers un modèle additif, une réponse s’exprime comme une somme de fonctions, chacune ayant un prédicteur pour argument. Pour cette raison, ces fonctions appelées fonctions coordonnées, sont des indicateurs de l’influence de chaque variable explicative sur la réponse. Ainsi, dans le cas de p prédicteurs $\{x^1, \dots, x^p\}$, le modèle s’écrit

$$\hat{y} = \sum_{k=1}^p s_k(x^k),$$

les p fonctions splines coordonnées $\{s_k(\cdot)\}_k$ étant estimées par PLSS. Ce type de modèle ne peut capturer d’éventuelles interactions entre deux ou plusieurs prédicteurs, mais il présente l’avantage d’être un modèle non linéaire multivarié aisé à interpréter. L’accent est mis tout au long de ce guide sur le rôle des super-paramètres qui contrôlent le modèle PLSS.

L’auteur tient particulièrement à remercier Stéphanie Bettan et le TEPRAL-centre de recherche branche boissons du groupe Danone, qui ont permis la diffusion de l’analyse sensorielle des données sur des jus d’orange traitées dans ce guide d’utilisation de PLSS. Que Rodolphe Seibt, ancien étudiant du DESS-MSIAAP de l’Université Montpellier II, soit aussi remercié pour sa collaboration dans la consultation qui a fait l’objet du traitement de ces données par PLSS.

2 Les splines de régression sous S-Plus

L'utilisation par les statisticiens des fonctions splines c'est à dire des polynômes par morceaux se raccordant en des points appelés noeuds, est de plus en plus courante que ce soit en analyse exploratoire des données ou en modélisation non-paramétrique.

Voici rapidement énumérés les premiers avantages des transformations splines.

- **Un espace vectoriel fonctionnel de dimension finie**

Un des éléments de leurs succès réside dans le fait qu'une spline est un élément d'un espace vectoriel fonctionnel de dimension finie. Ainsi que pour les polynômes usuels, une fois connue une base $\{B_1(\cdot), \dots, B_r(\cdot)\}$, une fonction spline comme celle représentée sur la Figure 0, est combinaison linéaire des fonctions de base

$$s(x) = \sum_{j=1}^r \beta_j B_j(x). \quad (1)$$

Les coordonnées $\beta = (\beta_1, \dots, \beta_r)$ de s dans la base, appelées coefficients splines, sont à estimer à partir d'un échantillon de la variable x sur un intervalle $[a, b]$ et d'un échantillon de la variable réponse y .

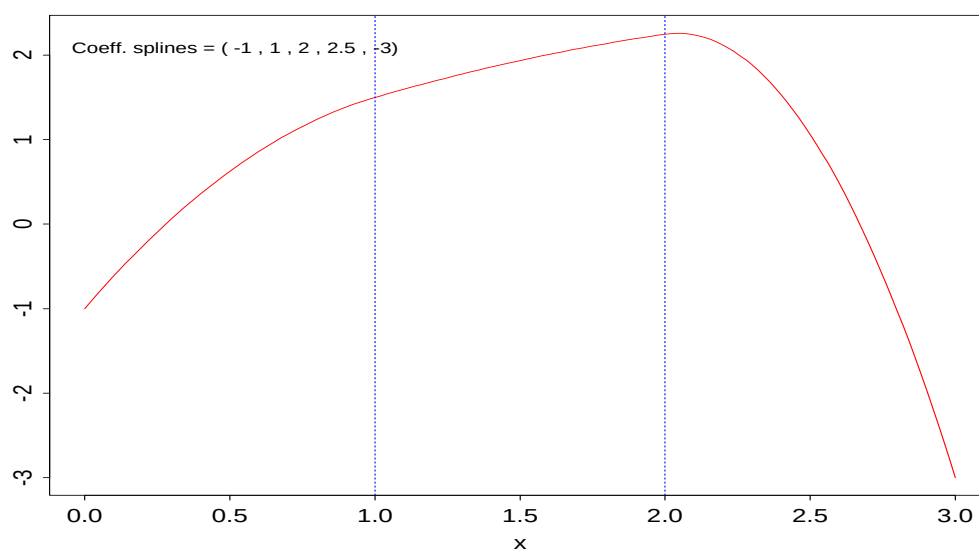


Figure 0: Fonction spline $s(x)$ de degré 2 sur $[0, 3]$ avec les noeuds en 1 et 2.

- **Des fonctions de base à support local**

L'avantage par rapport aux polynômes usuels est qu'une fonction spline de base est à support local, c'est à dire est nulle en dehors d'un certain intervalle. Ainsi

une spline fournit un outil statistique assez robuste face au problème des valeurs extrêmes des variables explicatives puisqu'une telle observation n'influencera s que par les seules fonctions de base dont le support contient cette donnée extrême ou aberrante. Le revers de la médaille de cet avantage est que $s(x) = 0$ en dehors de l'intervalle $[a, b]$. Ainsi les splines ne peuvent être utilisées en prédiction que sur cet intervalle. Bien sûr, on peut prolonger s linéairement en dehors de $[a, b]$ et ainsi construire une *prédiction à court terme*.

D'autres propriétés utiles pour la régression sont présentées dans les paragraphes suivants. Le langage S-Plus dispose d'une fonction native appelée **bs** qui permet de calculer les valeurs des fonctions de bases le plus populaires appelées les fonctions *B-splines* (De Boor 1978).

2.1 La fonction bs sous S-Plus

Résumé des principales commandes :

```
>help(bs)
```

```
>B<-bs(x,knots=c(1,2,4),degree=2,intercept=T)
```

On a déjà indiqué qu'une spline définie par (1) sur un intervalle $[a, b]$ est une fonction polynomiale par morceaux dont les morceaux (des polynômes de degré d) se raccordent ainsi que certaines de leurs dérivées en des points appelés *noeuds*.

L'espace vectoriel des fonctions splines d'ordre $m = d + 1$ sur un intervalle $[a, b]$ est parfaitement défini lorsque l'on a choisi

- le nombre K de noeuds intérieurs à $[a, b]$,
- la position des noeuds intérieurs notés $\{\tau_{m+1}, \dots, \tau_{m+K}\}$ car m noeuds confondus sont automatiquement placés en a et b . On dit que a et b sont des noeuds multiples de multiplicité m .

$$\tau_1 = \dots = \tau_m = a < \tau_{m+1} \leq \dots \leq \tau_{m+K} < b = \tau_{m+K+1} = \dots = \tau_{2m+K}.$$

On verra par la suite comment la multiplicité d'un noeud contrôle le nombre de dérivées qui se raccordent en ce point et donc la régularité locale de la spline s .

On note cet espace vectoriel $S(m, \{\tau_{m+1}, \dots, \tau_{m+K}\}, [a, b])$ qui est de dimension

$$r = m + K.$$

Une fois choisis les super-paramètres m et $\{\tau_{m+1}, \dots, \tau_{m+K}\}$, c'est à dire une fois défini l'espace vectoriel des splines, la fonction **bs** native sous S-Plus permet de calculer les valeurs des fonctions de base appelées des *B-splines* (*B* pour *Basis*), sur un échantillon de la variable x noté **x**. L'Annexe 3 présente l'expression mathématique des fonctions *B-splines* ainsi que la formule de récurrence permettant le calcul.

```
>help(bs)
```

Les arguments de cette fonction seront précisés au fur et à mesure de leur utilisation. L'un d'entre-eux, **df** pour "degree of freedom", est optionnel et correspond à la dimension r de l'espace spline lorsque **intercept=T**. Lorsque **df** est spécifié plutôt que **knots** qui est le vecteur des noeuds intérieurs, alors la fonction **bs** choisit **df + !intercept -degree -1** noeuds aux quantiles correspondants de **x** (**!intercept** est la négation d'**intercept** qui vaut 0 si **intercept=T** et 1 sinon). Dans la suite on spécifiera plutôt le degré et le nombre de noeuds intérieurs.

2.1.1 Le graphe des fonctions de base

Voici une première application de **bs** qui permet de tracer le graphe des fonctions *B-splines* pour $S(3, \{1, 2, 4\}, [0, 5])$ de dimension $r = 6$. Pour cela il suffit d'abord de fabriquer une grille de points **x** sur $[0, 5]$ suffisamment fine, par exemple 50 points équidistants,

```
>x<-seq(0,5,length=50)
```

puis, de construire la matrice **B** de dimensions 50×6 dont les colonnes donnent les valeurs des fonctions de base calculées sur **x**.

```
>B<-bs(x,knots=c(1,2,4),degree=2,intercept=T)
>par(mfrow=c(2,3))
>for(i in 1:6)
>+ plot(x,B[,i],xlab="",ylab="",ylim=c(0,1),main=paste("B",i),type="l")
```

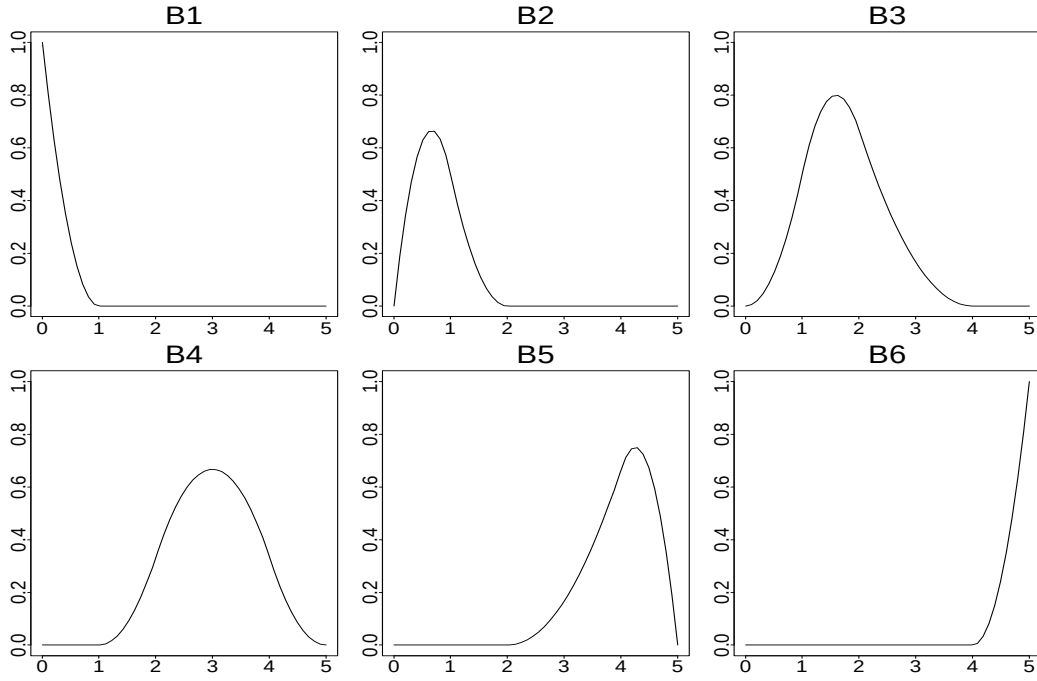


Figure 1: Fonctions B -splines pour l'espace spline $S(3, \{1, 2, 4\}, [0, 5])$.

La Figure 1 permet de visualiser certaines propriétés de la base $\{B_1, \dots, B_{m+K}\}$.

2.1.2 Propriétés des B -splines

P1. $0 \leq B_i(x) \leq 1$.

P2. **Une B -spline $B_i(\cdot)$ a pour support** $[\tau_i, \tau_{i+m}]$: $B_i(x) = 0$ pour $x \notin [\tau_i, \tau_{i+m}]$

Sur l'exemple, $B_1(x) > 0$ sur $[\tau_1 = 0, \tau_4 = 1[$, $B_2(x) > 0$ sur $]\tau_2 = 0, \tau_5 = 2[$

$B_3(x) > 0$ sur $]\tau_3 = 0, \tau_6 = 4[$, $B_4(x) > 0$ sur $]\tau_4 = 1, \tau_7 = 5[$

$B_5(x) > 0$ sur $]\tau_5 = 2, \tau_8 = 5[$, $B_6(x) > 0$ sur $]\tau_6 = 4, \tau_9 = 5]$

P3. Pour $x \in [a, b]$, il ne peut y avoir plus de m B -splines telles que $B_i(x) > 0$.

Sur l'exemple, $B_4(4.5) > 0$, $B_5(4.5) > 0$, $B_6(4.5) > 0$ alors que les autres fonctions donnent 0. Cette propriété précise l'influence des données extrêmes sur $s(x)$.

P4. **Régularité d'une spline $s(x)$ en un noeud τ_i de multiplicité m_i :**

Soit m_i la multiplicité du noeud τ_i telle que $1 \leq m_i \leq m$, alors les $m - 1 - m_i$ premières dérivées à gauche et à droite se raccordent au noeud τ_i

$$s_-^{(j)}(\tau_i) = s_+^{(j)}(\tau_i) \quad \text{pour} \quad j = 1, \dots, m - 1 - m_i.$$

Cette propriété est vraie aussi pour une spline de base $B_i(\cdot)$.

Sur l'exemple, $m = 3$ (splines de degré 2), les noeuds intérieurs 1, 2 et 4 sont simples ou de multiplicité $m_i = 1$, il en résulte que

les fonction B_1, B_2, B_3 et B_4 sont localement de classe C^1 autour de 1,
les fonctions B_2, B_3, B_4 et B_5 sont localement de classe C^1 autour de 2,
les fonctions B_3, B_4, B_5 et B_6 sont localement de classe C^1 autour de 4.

Pour les noeuds 0 et 5 de multiplicité $m_i = 3$, il faut prendre en compte les noeuds appartenant au support de la B -spline considérée:

B_1 de support $[\tau_1 = \tau_2 = \tau_3 = 0, \tau_4 = 1]$ est C^{-1} (discontinue) en 0.

B_2 de support $[\tau_2 = \tau_3 = 0, \tau_5 = 2]$ est C^0 (continue) en 0.

B_3 de support $[\tau_3 = 0, \tau_6 = 4]$ est localement C^1 en 0.

Même raisonnement, pour B_4, B_5 et B_6 au noeud 5.

Le plus souvent on utilise des splines à noeuds intérieurs simples (multiplicité 1) ce qui fournit la classe de splines la plus lisse (C^{m-2}) sur $]a, b[$. Bien sûr, $s(x)$ est toujours discontinue en a et b .

La propriété suivante non directement lisible sur la Figure 1, explique aussi l'intérêt porté à l'utilisation de la base des B -splines.

P5. Partition de l'unité et matrice de codage B :

$$\sum_{i=1}^r B_i(x) = 1, \text{ pour } x \in [a, b].$$

Cette propriété jointe à P1, est attractive si l'on désire obtenir un "codage flou" d'une variable. Pour cette raison on désigne parfois la matrice B comme la matrice de codage de l'échantillon $\mathbf{x}$ qui vérifie la relation

$$B\mathbf{I}_r = \mathbf{I}_n \quad (2)$$

où $\mathbf{I}_k$ désigne le vecteur colonne de 1 en nombre k .

```
>x[3]
>0.2040816
>B[3,]
      1      2      3 4 5 6
0.633486 0.3456893 0.02082466 0 0 0
>sum(B[3,])
[1] 1
```

Le centrage des colonnes de la matrice de codage a donc pour conséquence, comme

en Analyse Factorielle des Correspondances, de diminuer de 1 le rang de la matrice. On peut comprendre maintenant le rôle de l'argument booléen `intercept` de la fonction `bs` : la valeur par défaut `F` ne fournit en sortie que les $r - 1$ dernières colonnes de la matrice. Pour obtenir toutes les colonnes on spécifie `intercept=T`.

```
>B1<-bs(x,knots=c(1,2,4),degree=2)
>B1[3,]
      1      2 3 4 5
0.3456893 0.02082466 0 0 0
```

Lorsque `df` est spécifié la matrice de codage `B` possède `df` colonnes. La valeur par défaut `intercept=F` permet d'obtenir une matrice de codage centrée de plein rang colonne $r - 1$ (si le rang de `B` avant centrage est r) pour l'utiliser comme matrice "design" en régression linéaire multiple et ainsi obtenir un estimateur spline aux moindres carrés présenté dans la Section 2.2.

2.1.3 Exercices sous S-Plus

Exercice :

Ecrire les instructions S-Plus permettant d'obtenir les Figures 2 à 4.

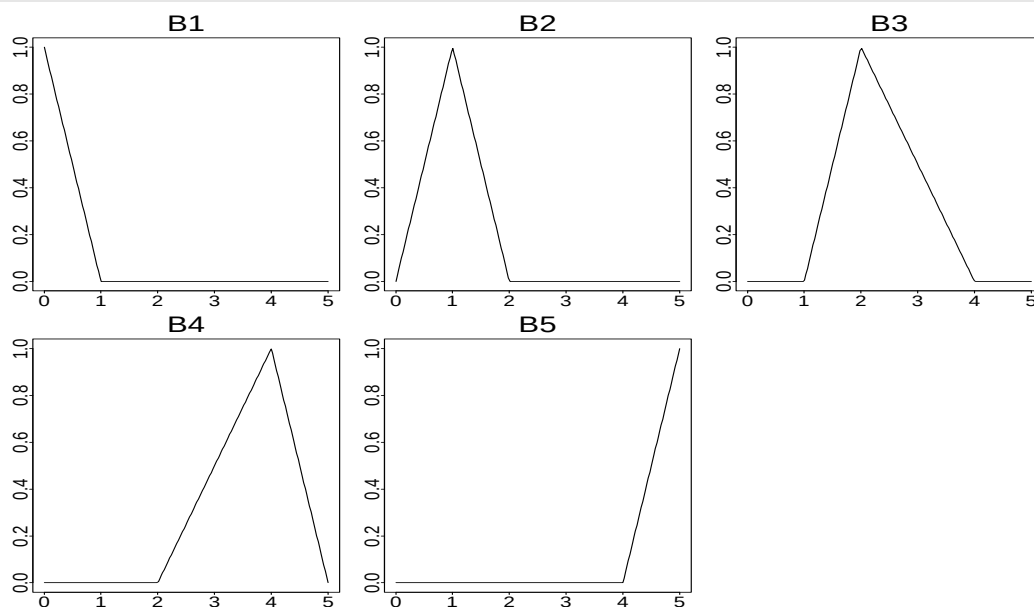


Figure 2: Fonctions B -splines pour l'espace spline $S(2, \{1, 2, 4\}, [0, 5])$.

La Figure 2 fournit une base pour les splines linéaires par morceaux. La Figure 3 montre des splines de base de degré 0 (constantes par morceaux).

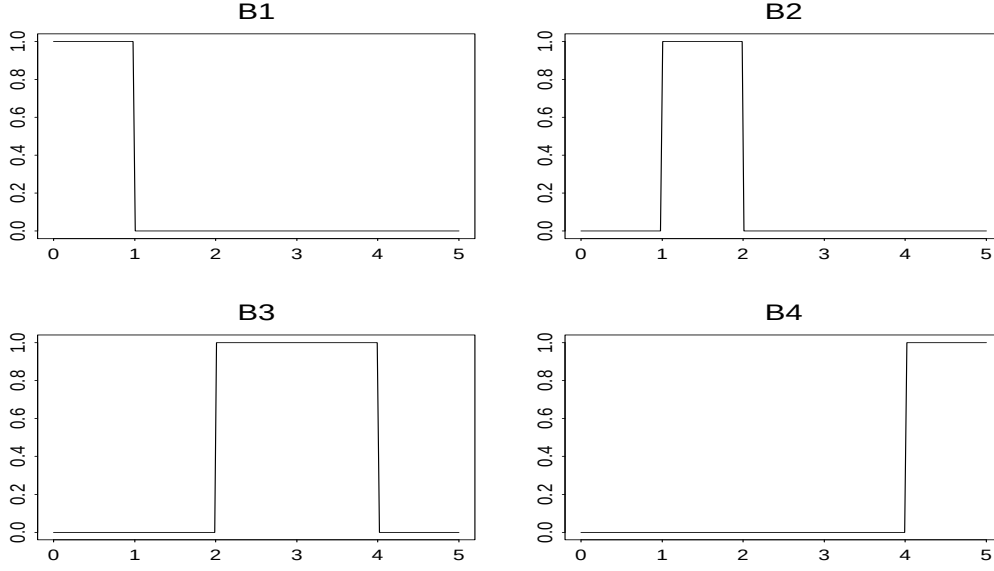


Figure 3: Fonctions B -splines pour l'espace spline $S(1, \{1, 2, 4\}, [0, 5])$.

Qu'arrive-t-il si aucun noeud intérieur n'est utilisé? La base obtenue est une base de l'espace des polynômes d'ordre m sur l'intervalle $[a, b]$.

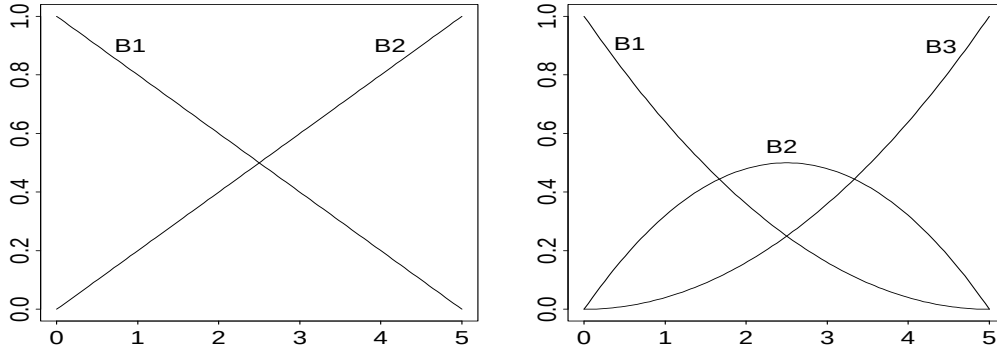


Figure 4: Fonctions B -splines de base pour les espaces $S(2, \emptyset, [0, 5])$ et $S(3, \emptyset, [0, 5])$ qui sont les espaces de polynômes de degré 1 et de degré 2 sur $[0, 5]$.

Cette propriété est importante du point de vue des applications. Les polynômes d'ordre m sur un intervalle $[a, b]$ sont des splines particulières d'ordre m lorsque aucun noeud intérieur à $[a, b]$ n'est utilisé. Tout modèle de régression utilisant une base de B -splines et en particulier PLSS, est capable de capturer les modèles linéaires et polynomiaux sur $[a, b]$ en ne prenant aucun noeud intérieur.

2.2 Les splines aux moindres carrés en régression simple

Résumé des principales commandes :

```
>bivarplot(juicX,juicY[,1,drop=F],matrow=3,matcol=4)
>help(lm)
>Bc<-bs(juicX[,1],degree=2,knots=c(167.4,257.2,415.8))
>summary(lm(juicY[,1]~Bc))
>plot(lm(juicY[,1]~Bc))
>resul<-lss(juicX[,1,drop=F],juicY[,1,drop=F],degree=2,knots=3)
```

Les observations d'une variable aléatoire continue y sont mesurées sur n valeurs d'une variable explicative x . On note X et Y les vecteurs des n observations sur le prédicteur et sur la réponse. Une fois choisi l'espace spline, c'est à dire le degré, le nombre et la position des noeuds intérieurs à l'étendue des valeurs de X , le modèle s'écrit

$$y_i = s(x_i) + \varepsilon_i \quad i = 1, \dots, n$$

et l'on suppose que les variables aléatoires ε_i sont indépendantes et suivent une même loi normale de moyenne 0 et de variance σ^2 . L'estimation des coefficients splines $\beta = (\beta_1, \dots, \beta_r)$ de (1) par les moindres carrés usuels OLS (Ordinary Least-Squares), conduit à l'estimateur spline LSS (Least-Squares Splines) ce que l'on résume par

$$LSS(X, Y) \equiv OLS(B, Y).$$

La matrice B est la matrice de codage des prédicteurs. L'estimateur LSS fait partie de la famille des lisseurs linéaires puisque $\hat{Y}$ qui approxime Y , vérifie

$$\hat{Y} = B\hat{\beta} = HY$$

où H , matrice de la projection orthogonale sur l'espace des colonnes de B encore appelée matrice de lissage, ne dépend pas de Y . La fonction S-Plus `lm` est l'outil de modélisation linéaire qui permet cette estimation. Cependant, les coefficients du modèle ne sont pas directement interprétables et les sorties de `lm` ne fournissent pas le graphe de la spline estimée. Aussi, la fonction S-Plus `lss` a-t'elle été construite par l'auteur pour visualiser graphiquement l'estimateur LSS. Si cet estimateur ne correspond pas à une bonne qualité d'ajustement, les graphes fournis par `lss` permettront d'explorer d'autres jeux de super-paramètres (position des noeuds intérieurs...).

2.2.1 Les données des jus d'orange

Les données qui sont présentées dans l'Annexe 1 sont issues d'une étude réalisée par R. Seibt et J.F. Durand pour le compte du centre de recherche TEPRAL du groupe Danone. Sur 24 jus d'orange ayant tous la même teneur en fruit et la même quantité de glucides, ont été mesurées 10 variables minéralogiques et 5 variables sensorielles. Pour des raisons de confidentialité, les individus "jus d'orange" sont repérés par les lettres capitales "A", ..., "X". Les 10 variables explicatives sont la **Conductivité** (*COND*), les 8 **paramètres minéralogiques** *SiO₂*, *Na*, *K*, *Ca*, *Mg*, *Cl*, *SO₄*, *HCO₃* et la **Somme** des éléments minéraux (*Somme*). Les 5 réponses sont les descripteurs sensoriels **Chargé** (*Char*), **Intensité** (*Int₁* et *Int₂*) et **Agressivité** (*Agr₁* et *Agr₂*) dont les protocoles de mesures ne sont pas révélés pour cause de confidentialité. L'objectif est l'étude des liens linéaires ou non linéaires entre les variables sensorielles et les variables analytiques. L'examen des corrélations montre que les prédicteurs *COND*, *Ca*, *Mg*, *SO₄* et *Somme* sont très corrélés ainsi que *SiO₂* et *K* et, du côté des réponses, une très forte corrélation entre les 4 variables d'Intensité et d'Agressivité.

Nous supposerons que les matrices *juicX* et *juicY* des mesures sur les prédicteurs et les réponses présentées dans l'Annexe 1, sont à disposition sous S-Plus. Dans un premier temps, la fonction `bivarplot` permet d'explorer les graphes bivariés entre une réponse, ici *Char*, et les 10 prédicteurs.

```
>bivarplot(juicX,juicY[,1,drop=F],matrow=3,matcol=4)
```

La spécification `drop=F` dans la sélection de la colonne 1 de la matrice *juicY* permet de ne pas perdre la structure matricielle du vecteur colonne 1 et ainsi de conserver le nom de la variable *Char*. La matrice de 3 × 4 graphiques bivariés obtenus est présentée dans la Figure 5. Un lissage des observations est présenté pour chaque graphe bivarié. Les lisseurs de l'option `smooth` sont les lisseurs classiques disponibles sous S-Plus (Venables et Ripley 1994). La valeur par défaut est `smooth='lowess'`, les autres possibilités sont présentées dans l'Annexe 2.

Presque toutes les variables explicatives présentent des observations extrêmes qui vont avoir une forte influence sur les différents modèles examinés. La propriété P2 des fonctions *B-splines* permet par un choix judicieux des noeuds, que les différents modèles

splines examinés par la suite ne soient que localement influencés par ces données.

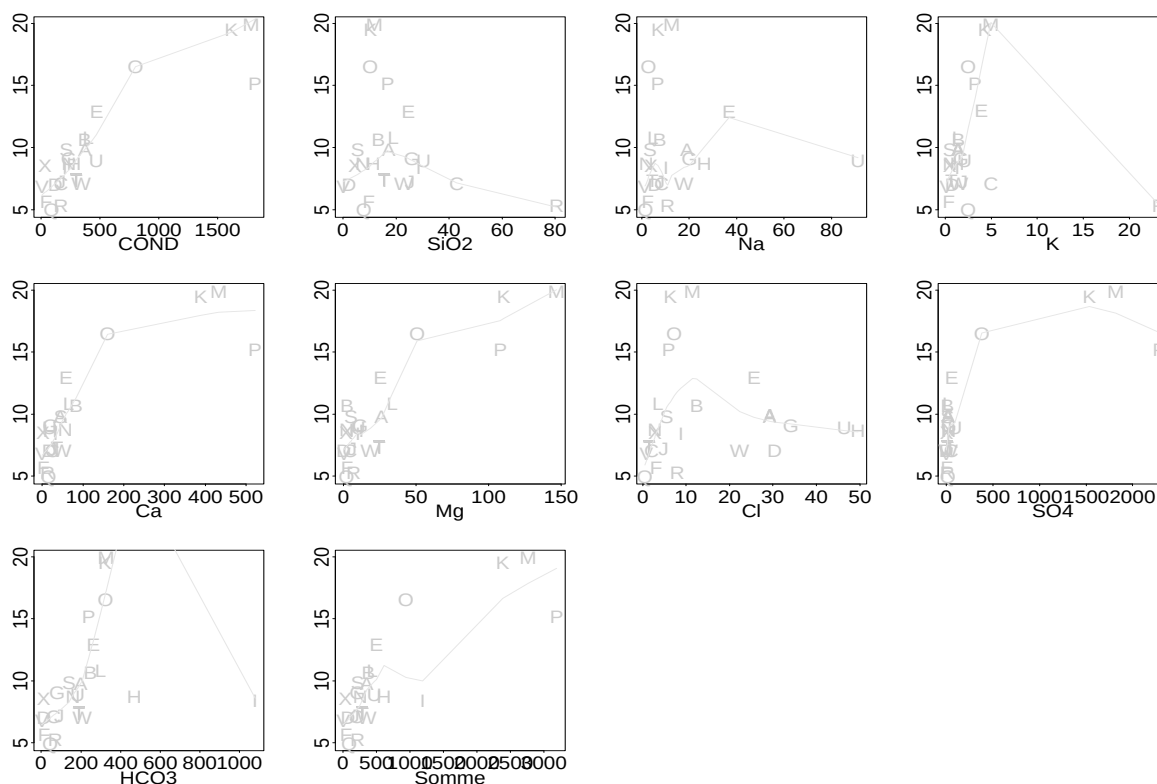


Figure 5: Graphes bivariés entre la réponse **Chargé** et chacun des prédicteurs. Le lissage des observations est obtenu par défaut par la fonction “lowess”.

2.2.2 La régression LSS entre Chargé et Conductivité

La fonction S-Plus native permettant la régression spline aux moindres carrés est la fonction `lm` couplée avec la fonction `bs`. Nous ne commenterons pas les graphiques obtenus par l’instruction

```
plot(lm(juicY[,1] ~ bs(juicX[,1],degree=2,knots=c(167.4,257.2,415.8))))
```

qui ne donne pas le graphe de l’estimateur spline. C’est pour cette raison que la fonction `lss` a été construite afin de visualiser le graphe de la fonction spline dans le cadre univarié et les graphes des fonctions splines coordonnées dans le cadre multivarié additif. Pour la régression simple entre *Char* et *COND* comparons les coefficients des modèles obtenus par `lm` et par `lss` lorsque le degré choisi est 2 et 3 noeuds sont placés aux quartiles de *COND*.

```
>help(lm)
>Bc<-bs(juicX[,1],degree=2,knots=c(167.4,257.2,415.8))
>summary(lm(juicY[,1]~Bc))
```

```
Call: lm(formula = Char ~ Bc)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-2.266 -0.7589  0.0509  1.165  2.16
```

```
Coefficients:
```

	Value	Std. Error	t value	Pr(> t)
(Intercept)	7.6654	1.1452	6.6934	0.0000
Bc1	-3.7547	2.3736	-1.5818	0.1311
Bc2	1.3844	1.2909	1.0724	0.2977
Bc3	0.6036	1.4134	0.4271	0.6744
Bc4	18.2951	3.2757	5.5851	0.0000
Bc5	9.6787	1.5186	6.3736	0.0000

```
Residual standard error: 1.455 on 18 degrees of freedom
```

```
Multiple R-Squared: 0.9032
```

```
F-statistic: 33.59 on 5 and 18 degrees of freedom, the p-value is 1.607e-08
```

```
Correlation of Coefficients:
```

	(Intercept)	Bc1	Bc2	Bc3	Bc4
Bc1	-0.7768				
Bc2	-0.7433	0.3546			
Bc3	-0.8427	0.7054	0.5184		
Bc4	-0.3209	0.2043	0.3346	-0.0059	
Bc5	-0.7602	0.6000	0.5448	0.6993	0.0400

Remarquons que les colonnes de la matrice Bc sont les 5 dernières colonnes de la matrice de codage B obtenue avec l'argument `Intercept = T`.

```
>B<-bs(juicX[,1],degree=2,knots=c(167.4,257.2,415.8),intercept=T)
```

Le modèle linéaire s'écrit matriciellement

$$\widehat{Char} = 7.6654 \mathbb{I}_{24} - 3.7547 Bc^1 + \dots + 9.6787 Bc^5.$$

En fait, la propriété (2) de la matrice de codage B s'écrit sur l'exemple

$$B\mathbf{I}_6 = \mathbf{I}_{24}$$

ce qui permet de récrire le modèle linéaire en les colonnes de B

$$\widehat{Char} = 7.6654B^1 + (7.6654 - 3.7547)B^2 + \dots + (7.6654 + 9.6787)B^6,$$

et d'obtenir d'une autre façon les coefficients splines de la spline aux moindres carrés

$$\hat{\beta}_{LSS} = (B'B)^{-1}B'Y. \quad (3)$$

La fonction `lss` permet d'une part d'obtenir ces coefficients grâce à (3) sous réserve que $B'B$ soit inversible et d'autre part de représenter le graphe de la spline aux moindres carrés. A la différence de la fonction native `bs`, le paramètre `knots` de `lss` donne le nombre et non la position des noeuds qui est spécifiée soit explicitement par `listknots`, soit par défaut aux quantiles. L'Annexe 2 détaille le rôle des paramètres d'entrée de la fonction `lss`.

```
>resul<-lss(juicX[,1,drop=F],juicY[,1,drop=F],degree=2,
>+ listknots=list(c(167.4,257.2,415.8)))
```

L'instruction précédente place 3 noeuds aux quartiles du prédicteur. Remarquons que l'on obtient les mêmes résultats par l'instruction

```
>resul<-lss(juicX[,1,drop=F],juicY[,1,drop=F],degree=2,knots=3)
```

car, joint au paramètre `knots`, le paramètre booléen `equiknots` par sa valeur par défaut `F`, place les noeuds aux quantiles de la (ou des) variable(s) explicative(s); la valeur `T` fournit les noeuds également espacés.

R2

```
[1] 0.9031894
```

```
Least-squares fit
```

```
[1] 1.586859
```

```
The B-spline transformations of the 1 predictor(s) are now being computed
```

```
1
```

La dernière information indique qu'une grille de points (par défaut `nbpoints = 50`) est transformée par la base de splines pour pouvoir représenter la fonction spline LSS sur l'étendue des valeurs de la variable explicative.

L'erreur quadratique calculée est la moyenne des carrés des erreurs. Le lien avec celle obtenue par la fonction `lm` est

$$MSE_{lm} = \sqrt{\frac{MSE_{lss} \times \text{nbr individus}}{\text{degrés de liberté}}}.$$

```
>sqrt( 1.586859 * 24 / 18)
[1] 1.454583
```

Les critères du R^2 et de l'erreur quadratique moyenne sont des indices de bon ajustement. La question cruciale est celle du sur-ajustement aux données et l'exercice proposé plus loin est une exploration de ce problème à l'aide de la fonction `lss`.

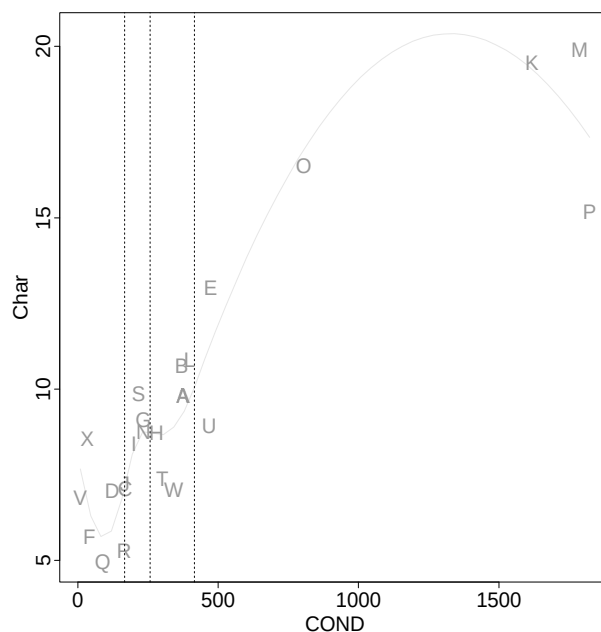


Figure 6: Fonction spline aux moindres carrés de degré 2. Les verticales indiquent la position des 3 noeuds.

Plots of the response residuals ? (y/n) 1: y

Plot of Yhat against Y (y/n)? 1: y

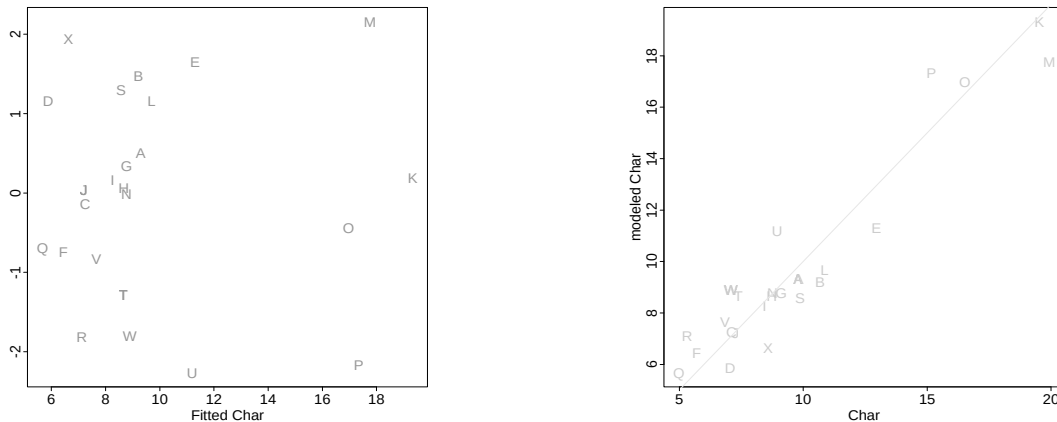


Figure 7: Graphique des résidus et graphique des couples $(Y, \hat{Y})$.

Ces deux derniers graphiques sont les mêmes que ceux obtenus grâce à la fonction `lm`.

On retrouve les coefficients (3) de la fonction spline aux moindres carrés par

```
>resul$beta
      [,1]
[1,] 7.665379
[2,] 3.910726
[3,] 9.049799
[4,] 8.268991
[5,] 25.960441
[6,] 17.344066
```

2.2.3 Exercices sous S-Plus

Exercices :

Faire différents essais de jeux de noeuds et de degrés pour la fonction `lss` appliquée à `Char` et `COND` avec comme objectif un équilibre entre un bon ajustement et une économie des super-paramètres (dimension d'espace spline).

L'option par défaut de `lss` permet des débuts raisonnables pour cette exploration.

```
>lss(juicX[,1,drop=F],juicY[,1,drop=F])
```

Par défaut, `degree=1` et `knots =0` ce qui code la (ou les) variable(s) explicative(s) par une (ou des) base(s) de splines de degré 1 avec 0 noeud. Le modèle obtenu est donc le modèle de la régression linéaire ordinaire simple (ou multiple). On peut ensuite continuer

en augmentant progressivement le degré tout en maintenant le nombre de noeuds égal à 0, ce qui permet d'explorer les modèles polynomiaux.

On peut alors poursuivre l'exploration en prenant un nombre de noeuds non nul et en spécifiant leurs positions. Des indications utiles à la recherche :

- degré 1 (splines linéaires) : placer les noeuds là où les données semblent indiquer un changement de pente.
- degré 2 (splines quadratiques) : choisir pour les noeuds les optimums locaux ou les points d'inflexion.

Il est clair que l'inflexion de la courbe de régression de la Figure 6, au voisinage de la médiane de *COND* est difficilement interprétable pour le praticien et qu'un modèle plus économique est à rechercher.

2.3 Le modèle spline additif dans le cadre multivarié

Considérons p prédicteurs $x^1, \dots, x^p$ et de q réponses $y^1, \dots, y^q$ mesurés sur un échantillon de n observations. Une fois choisis pour chaque prédicteur, les super-paramètres que sont le degré des polynômes, le nombre et la position des noeuds, le modèle spline additif s'écrit

$$y_i^j = s_1^j(x_i^1) + \dots + s_p^j(x_i^p) + \varepsilon_i^j, \quad j = 1, \dots, q, \quad i = 1 \dots, n, \quad (3)$$

où la k ième fonction spline coordonnée $s_k^j(\cdot)$, inconnue mais non-aléatoire, est censée exprimer l'influence du prédicteur x^k sur la réponse y^j . Les variables sont supposées centrées ainsi que les valeurs prises par chaque spline coordonnée. Les variables aléatoires $\{\varepsilon_i^j\}_i$ sont supposées suivre les mêmes hypothèses que dans le cadre de la régression simple.

Notons r_k la dimension d'espace spline pour la variable x^k et

$$r = \sum_{k=1}^p r_k.$$

La super-matrice de codage

$$B = [B^1 | \dots | B^p] \quad (4)$$

est la matrice de dimension $n \times r$ dont les blocs colonnes B^k sont les matrices de codage des prédicteurs par les bases de B -splines utilisées. Comme en Analyse Factorielle des

Correspondances Multiples, B n'est pas de plein rang colonne à cause de la propriété P5 des splines de base concernant les blocs B^k . De fait, le rang de B vérifie

$$\text{rang}(B) \leq \min(n, r - p + 1),$$

et si l'on centre les colonnes de B , alors

$$\text{rang}(B) \leq \min(n - 1, r - p).$$

Matriciellement, l'estimateur spline additif s'écrit

$$\hat{Y}^j = B\hat{\beta}^j = \sum_{k=1}^p B^k \hat{\beta}_k^j, \quad j = 1, \dots, q. \quad (5)$$

Plusieurs méthodes de construction des vecteurs $\hat{\beta}_k^j$ conduisent à différents estimateurs des fonctions splines coordonnées $s_k^j(x^k)$. Nous verrons rapidement les limites de l'estimateur aux moindres carrés usuel pour nous concentrer sur l'estimateur PLS.

Lorsque les deux critères de validité d'un modèle sont remplis, i.e. un bon ajustement aux données et une bonne qualité de prédiction, l'examen des pq courbes $s_k^j(x^k)$ est d'une aide précieuse pour interpréter le phénomène modélisé. Ceci n'est cependant envisageable que lorsque le nombre de variables est raisonnable. Dans certains problèmes comme celui de la modélisation de variables chimiques basée sur les spectres d'absorbance proches de l'infra-rouge (NIR), le nombre de prédicteurs est très grand (plusieurs centaines) et la visualisation des "courbes coordonnées" n'est envisageable qu'armé d'une grande patience... L'objectif de la modélisation est plutôt dans ce cas, de détecter les groupes prédicteurs, c'est à dire les zones spectrales, subissant le même type de transformation. On peut aussi sélectionner les variables explicatives significativement les plus influentes. Le modèle additif permet de classer les prédicteurs (préalablement centrés réduits) par ordre décroissant d'influence en prenant pour critère un paramètre de dispersion sur les valeurs des fonctions splines coordonnées calculées sur l'échantillon.

La façon la plus naturelle de construire un modèle spline additif est d'utiliser l'estimateur Least Square Splines (LSS) aux moindres carrés usuels qui a déjà été présenté dans le cadre de la régression simple

$$LSS(X, Y) = OLS(B, Y).$$

Les remarques faites sur le rang de B obligent d'utiliser l'option par défaut `intercept=F` de la fonction `bs` pour diminuer de 1 le nombre des colonnes des blocs B^k et ainsi rendre la matrice B centrée de plein rang colonne.

Cependant, sur les données des jus d'orange, on atteint vite les limites de cette approche qui sont celles de la régression linéaire multiple, puisque le prix à payer pour le passage du linéaire au non linéaire grâce aux splines de régression est "l'expansion de la dimension". Prenons l'exemple de la modélisation spline de **Chargé** par les 3 prédicteurs **Conductivité**, SiO_2 et Na .

```
>lm(juicY[,1]~bs(juicX[,1],degree=3,df=7)+bs(juicX[,2],degree=3,df=9)+
>+ bs(juicX[,3],degree=3,df=9))
```

```
Error in lm.fit.qr(x,y): computed fit is singular, rank 24
```

```
Dumped
```

Un message d'erreur apparaît.

Il provient du fait que les observations des 3 variables explicatives étant codées par des matrices B^1 , B^2 et B^3 de dimensions respectives 24×7 , 24×9 et 24×9 la fonction `lm` effectue une régression linéaire multiple basée sur 25 variables explicatives et 24 individus. La super-matrice de codage B a pour dimensions 24×25 et l'inversion de $B'B$ est impossible.

Dans le cadre multivarié, ce guide ne présente pas les modèles additifs généralisés obtenus par la fonction `gam` native sous S-plus et dont la référence est le livre de Hastie et Tibshirani (1990). La raison principale outre le manque de place, est que le présent guide trouve son unité dans le passage du linéaire au non linéaire par le simple codage des variables explicatives par une base de fonctions splines. D'autre part, la plupart des méthodes produisant un modèle additif sont sensibles au problème de la multicolinéarité c'est à dire à la présence de fortes corrélations entre certains prédicteurs, cause bien connue de détérioration de la qualité du modèle.

Les limitations précédentes plaident donc en faveur d'une méthode de modélisation comme PLSS qui d'une part supporte bien le passage du linéaire au non linéaire au prix de l'augmentation de la dimension et qui d'autre part apporte une réponse au problème de la multicolinéarité par une approche factorielle.

3 La régression PLS Spline (PLSS)

La régression PLS linéaire (Wold H. 1975) dont la version multi-réponses est due à Wold S., Martens et Wold H. (1983), construit une suite de variables explicatives non-corrélées notées $\{t^1, \dots, t^A\}$ et appelées variables latentes ou composantes principales, pour prédire les réponses de façon linéaire. Elle diffère principalement de la régression sur composantes principales (RCP) en ce que les composantes PLS sont construites en même temps que des régressions (partielles) sont effectuées. Cette simultanéité confère aux composantes PLS un meilleur pouvoir de prédiction que leurs homologues de la RCP. Voici résumées dans le paragraphe suivant les principales propriétés du modèle PLS linéaire. Nous renvoyons à (Durand, Roman et Vivien, 1998) pour plus de détails sur la construction du modèle PLS et sur l'utilisation de la fonction `pls` implémentée dans le langage S-Plus par Durand et Sabatier.

3.1 Le modèle PLS (linéaire)

Les colonnes des matrices $E_0 = X$ et $F_0 = Y$ des échantillons sur les prédicteurs et les réponses sont supposées centrées-réduites. L'étape k de PLS qui consiste à construire la k ème composante PLS t^k , se décompose en deux parties

1) Construction de t^k , à E_{k-1} et F_{k-1} fixés	$t = E_{k-1}w, \quad u = F_{k-1}c$ $(w^k, c^k) = \arg \max_{w'w=c'c=1} cov(t, u)$ $t^k = E_{k-1}w^k, \quad u^k = F_{k-1}c^k$
2) Actualisation de E_k et de F_k	$E_k = E_{k-1} - P_{t^k} E_{k-1}$ $F_k = F_{k-1} - P_{t^k} F_{k-1}$

La deuxième partie consiste à actualiser les prédicteurs et les réponses comme résidus de la régression sur la composante PLS calculée dans la première partie. La matrice P_{t^k} étant la matrice (de rang 1) de la projection orthogonale sur la composante t^k . Ces deux étapes sont imbriquées lorsque l'algorithme NIPALS dû à Wold (1966), est utilisé pour le traitement des données manquantes. Les fonctions `pls` et `plss` implémentées sous S-Plus ne prennent pas en compte, en l'état actuel de leur développement, le traitement

des données manquantes, voir Tenenhaus (1998) pour NIPALS dans la présentation de PLS.

Dans la première partie, la fonction objectif à maximiser est la covariance entre les compromis linéaires t et u . PLS diffère donc de l'analyse des corrélations canoniques par le critère (covariance versus corrélation) mais aussi par le traitement dissymétrique des régressions dans la partie 2 : les prédicteurs et les réponses actualisés sont régressés sur la composante “du côté des X ”, à savoir t^k . Il est à noter que dans le cas de la régression univariée, la variable u^1 est confondue avec la variable réponse Y . Le triplet solution $(cov(t^k, u^k), w^k, c^k)$ est celui associé à la plus grande valeur singulière, $cov(t^k, u^k)$, dans la décomposition en valeurs singulières de la matrice $\frac{1}{n}E'_{k-1}F_{k-1}$ des covariances géométriques “entre” les variables explicatives et les réponses actualisées.

Les premiers commentaires précédents éclairent de façon particulière une des propriétés de PLS comparée à l'Analyse en Composantes Principales (ACP) qui peut être considérée comme une “auto-PLS” de X sur lui même

$$PLS(X, Y = X) = ACP(X),$$

dans ce cas, $t^k = u^k$ et le critère est celui de la variance maximale.

Les composantes $\{t^1, \dots, t^A\}$ sont des variables centrées, deux à deux non-corrélées qui forment une base orthogonale d'un sous-espace vectoriel de $Im X$, espace engendré par les colonnes de X

$$t^k = X\alpha^k, \quad k = 1, \dots, A. \quad (6)$$

Notons $T_A = [t^1, \dots, t^A]$ la matrice dont les colonnes sont les composantes PLS. Le modèle PLS linéaire en les composantes, permet de reconstruire la matrice X et d'expliquer Y

$$\begin{aligned} X &= \hat{X}_A + E_A = P_{T_A}X + E_A \\ Y &= \hat{Y}_A + F_A = P_{T_A}Y + F_A. \end{aligned} \quad (7)$$

La matrice P_{T_A} , de rang A , est la matrice de projection orthogonale sur $Im T_A$.

Lorsque le nombre A de composantes est égal au rang de X , alors $Im T_A = Im X$ ($E_A = 0$) et le modèle PLS s'identifie au modèle de la régression linéaire multiple,

$$PLS(X, Y) = OLS(X, Y) \quad \text{si} \quad A = rang(X). \quad (8)$$

Les équations (6) et (7) permettent d'écrire le modèle PLS en les variables explicatives

$$\hat{Y}_A^j = X\hat{\beta}^j(A), \quad j = 1, \dots, q, \quad (9)$$

où les paramètres $\hat{\beta}$ dépendent du nombre A de composantes qui est le super-paramètre de la méthode. On exprime cela en notant $\hat{\beta}^j(A)$ le vecteur des paramètres modélisant la j ième réponse. Le choix de A se fait de façon usuelle par validation croisée.

3.2 Le modèle PLSS

L'estimateur spline additif PLSS (Durand 1997,1999) est obtenu par la régression PLS de Y sur le codage B des prédicteurs défini par (4). La méthode PLSS se résume par

$$PLSS(X, Y) = PLS(B, Y). \quad (10)$$

Les colonnes des matrices X et Y sont habituellement standardisées mais les colonnes de B sont seulement centrées. D'autre part, le fait que la super-matrice de codage B ne soit pas de plein rang colonne est seulement ici un artefact algébrique de peu d'importance statistique puisque les régressions partielles sont effectuées sur des composantes principales non corrélées issues de B . En outre, la propriété (8) du modèle PLS linéaire permet d'établir le lien entre PLSS et l'estimateur LSS des splines aux moindres carrés lorsque ce dernier est utilisable,

$$PLSS(X, Y) = OLS(B, Y) = LSS(X, Y) \quad \text{si} \quad A = rang(B).$$

Comme illustration de cette propriété, les Figures 6 et 8 représentant la régression de $Char$ sur $COND$ respectivement par LSS et PLSS, sont identiques.

```
>plss.plot(plss(juicX[,1,drop=F],juicY[,1,drop=F],standX=F,standY=F,
> + degree=2,knots=3,A=5))
```


Predictors' influence on Char (5 dim.)

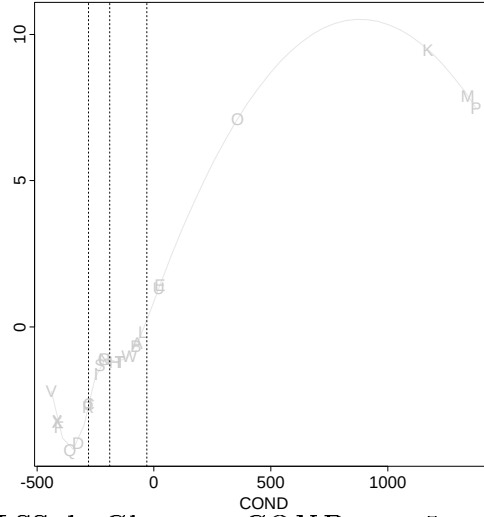


Figure 8: Modélisation PLSS de *Char* par *COND* avec 5 composantes et des splines de degré 2 avec 3 noeuds aux quartiles.

L'espace spline utilisé est le même. Cinq composantes PLSS permettent de reconstruire la matrice B de dimensions 24×6 , mais de rang 5 à cause du centrage. Le coefficient $R^2 = 0.903$ est identique à celui obtenu par LSS qui se confond avec PLSS lorsque $A = \text{rang}(B)$. Sur la Figure 8, les variables sont centrées et les observations du couple $(COND, Char)$ ne sont pas représentées, seuls sont représentés les couples $(COND, \widehat{Char})$, ce qui induit une échelle sur l'axe vertical différente de celle de la Figure 6.

Les composantes t^k qui dans PLS étaient grâce à (6), des compromis linéaires des prédicteurs, s'expriment dans PLSS comme des compromis des colonnes de B , c'est à dire par regroupement d'indices, comme la somme de p splines coordonnées

$$t^k = B\alpha^k = \sum_{i=1}^p B^i \alpha_i^k, \quad k = 1, \dots, A. \quad (11)$$

L'examen des graphes des fonctions coordonnées permet d'interpréter les composantes PLSS en fonction des prédicteurs. Les utilisateurs des méthodes factorielles linéaires, habitués à interpréter les plans factoriels des individus par l'examen des cercles des corrélations, devront se familiariser avec la compréhension de ces composantes qui capturent des non-linéarités sur les prédicteurs.

Le modèle spline additif (5) associé à PLSS qui est la transcription du modèle PLS (9) via la définition (10), est obtenu par regroupement des coefficients

$$\widehat{Y}_A^j = B\widehat{\beta}^j(A) = \sum_{i=1}^p B^i \widehat{\beta}_i^j(A), \quad j = 1, \dots, q. \quad (12)$$

Là aussi comme pour les composantes, le modèle (12) qui capture l'influence non linéaire additive des prédicteurs, s'interprète par l'examen des fonctions coordonnées les plus significatives. On verra dans le paragraphe suivant comment les p prédicteurs sont classés par influence décroissante selon un critère de dispersion de $B^i \widehat{\beta}_i^j(A)$ qui est l'étendue des valeurs d'une variable transformée.

Pour terminer cette introduction à PLSS, rappelons les super-paramètres du modèle :

- le type d'espace spline utilisé pour les fonctions splines coordonnées (degré des polynômes, nombre et position des noeuds),
- le nombre de composantes principales.

4 Pratique de PLSS

Comme `pls`, la fonction `plss` dispose d'une interface conversationnelle. La décision qui incombe à l'utilisateur est le choix des espaces splines pour les prédicteurs. Par défaut, `plss` propose l'option des splines de degré 1 avec aucun noeud qui, par l'utilisation de bases polynomiales de degré 1, construit un modèle PLS linéaire. La prise en compte d'éventuelles non-linéarités dépendra de la stratégie de l'utilisateur quant à l'augmentation progressive du degré et des noeuds. Une fois arrêté le choix des splines, l'interface conversationnelle des fonctions `plss`, `plss.plot`, `plssc` et `plssc.plot`, répond à différents objectifs.

- **Evaluation de la qualité de l'ajustement** par les coefficients R^2 des réponses avec les sous-espaces principaux emboîtés $\{t^1, \dots, t^A\}$.
- **Exploration des données** par l'interprétation des plan principaux (t^i, t^j) .
- **Interprétation des modèles additifs** pour les composantes $\{t^k\}$ et les réponses.
- **Validation du nombre de composantes** par une étape de cross-validation.
- **Prédiction et/ou validation externe** sur un éventuel deuxième jeu de données.

L'information acquise lors des premiers traitements du même jeu de données permettra ensuite d'essayer des placements de noeuds individualisés pour chacune des variables

explicatives, si celles-ci ne sont pas trop nombreuses... Les exemples présentés dans les paragraphes suivants illustrent séparément un ou plusieurs des points précédents, le soin est laissé au lecteur-utilisateur du logiciel, de parcourir toutes les sorties numériques et graphiques et d'interpréter les items non commentés dans ce guide.

L'examen des corrélations entre les réponses montre que la variable **Chargé** est peu corrélée avec le groupe des variables d'intensité et d'agressivité qui sont par contre très fortement mutuellement corrélées. L'étude des liens linéaires ou non entre les variables sensorielles et les variables minéralogiques nécessite deux régressions PLSS distinctes, l'une consistant à modéliser **Chargé**, l'autre à modéliser simultanément le groupe des 4 variables d'intensité de d'agressivité.

4.1 La régression de Chargé

Résumé des principales commandes

```
>essai0<-plss(juicX,juicY[,1,drop=F])
>plss.plot(essai0)
>essai0cv<-plsscvcv(juicX,juicY[,1,drop=F],Y,A=9,prop=0.05)
>plscvcv.plot(essai0cv)
>essai1<-plss(juicX,juicY[,1,drop=F],degree=2,listknots=localnoeuds)
```

4.1.1 L'option par défaut de PLSS : PLS linéaire

Dans la version la plus simple de **plss**, les bases de splines utilisées sont du même type pour toutes les variables : degré 1 (**degree=1**), nombre de noeuds intérieurs égal à 0 (**knots=0**), ce qui fournit une base de polynômes de degré 1 pour chaque prédicteur. Ainsi, l'option par défaut de **plss** est la régression PLS sur un codage linéaire des prédicteurs. Cependant, PLSS linéaire diffère de PLS standard au sens où les coefficients du modèle sont associés aux fonctions linéaires de base dont le graphe est présenté sur la Figure 4, et non directement aux variables explicatives.

```
>essai0<-plss(juicX,juicY[,1,drop=F],A=1)
```

PLSS : PLS LINEAIRE SUR LE CODAGE SPLINE DES PREDICTEURS

- PLS lineaire -

Variance totale de Y = 1

Variance totale de B = 1.238064

La variance de la réponse standardisée est 1, alors que celle des 10×2 colonnes centrées de B , est 1.24. Cette variance est plus faible que la variance totale des 10 variables explicatives normalisées, ce qui induit des composantes de variance plus faible que celles de PLS standard.

Dimension 1

cov(t,u)=0.738041 r(t,u)=0.868 stdev(t)=0.8501486 stdev(u)= 1

Char % VAR Y

R2 part. 0.754 75.365

% VAR B expliquée par la comp. = 58.85448

nombre d'axes supplémentaires ?1: 0

Par défaut, le nombre des composantes c'est à dire la dimension du modèle, est $A=2$. La matrice B de dimensions 24×20 qui devrait être de rang 10 à cause du centrage, est de rang 9 puisque la variable **Somme** est une combinaison linéaire des autres. La dimension maximale du modèle PLSS linéaire est $A = 9$. Remarquons au passage que la régression linéaire multiple n'est pas ici envisageable pour cette raison.

Le choix du nombre de composantes est crucial pour PLS. Les valeurs successives des coefficients R^2 avec les sous-espaces vectoriels emboîtés engendrés par les composantes, les " R^2 cumulés", forment une suite croissante qui fournit un critère d'ajustement aux données. Ici on choisit une seule composante puis on met en oeuvre une étape de validation croisée dont le critère ou PRESS basé sur la prédiction "à l'intérieur des données", fournit un autre indicateur pour le choix de la dimension du modèle.

```
>essai0cv<-plsscvcv(juicX,juicY[,1,drop=F],A=9,prop=0.05)
>plscvcv.plot(essai0cv)
```

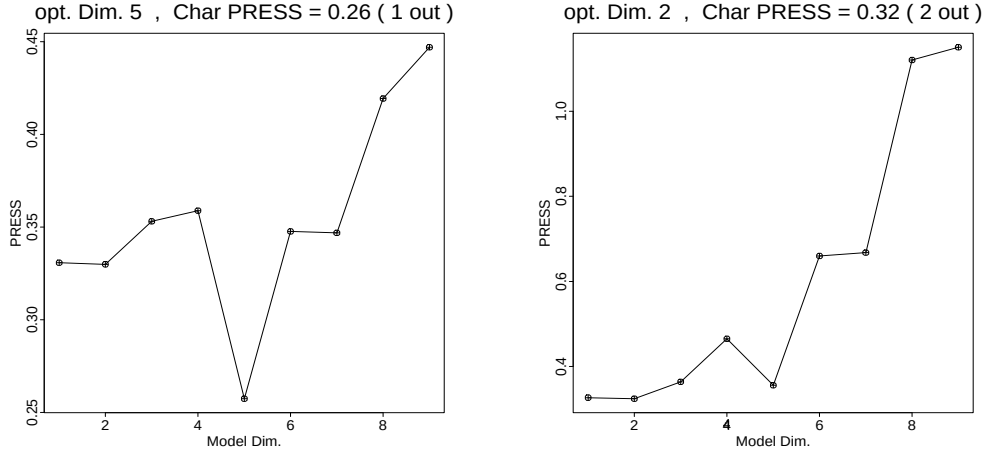


Figure 9: Différents PRESS du modèle PLSS linéaire **Chargé** en fonction de la dimension selon que un seul (**prop**=0.05) ou deux individus (**prop**=0.1) sont enlevés à la fois.

Après avoir enlevé un individu i (ou la proportion 10% des individus, ce qui est l’option par défaut **prop**=0.1), c’est à dire une ligne des matrices B et Y , on calcule la matrice $\hat{\beta}^{(-i)}(A)$ des coefficients du modèle PLSS à A dimensions, construite avec les individus restants. La répétition de cette opération pour $i = 1, \dots, n$, permet de calculer la moyenne des carrés des erreurs de prédiction ou $PRESS^j(A)$, pour chacune des réponses Y^j . Dans le cadre multi-réponses ($q > 1$), le PRESS total est la somme

$$PRESS(A) = \sum_{j=1}^q PRESS^j(A).$$

La Figure 9 montre deux résultats différents du choix de A selon que l’on enlève un seul individu (**prop**=0.05) ou deux individus à la fois (**prop**=0.1). Dans le premier cas, la dimension A_{opt} qui réalise le minimum des valeurs $\{PRESS(A)\}$ est égale à 5 avec $PRESS(A_{opt}) = 0.26$, dans le second $A_{opt} = 2$ avec $PRESS(A_{opt}) = 0.32$. En fait lorsqu’un groupe d’individus est enlevé, le résultat est ici fortement dépendant de l’ordre dans lequel les individus sont enlevés et prédits. En effet, certains jus d’orange sont fortement atypiques sur certaines variables et il est recommandé de mettre en oeuvre un certain nombre d’étapes de validation croisée sur différentes permutations des individus (utiliser la fonction S-Plus native **sample** pour générer des échantillons permutés).

Pour décider de la dimension du modèle deux considérations sont à prendre en compte: la qualité de prédiction “interne” (évolution de $PRESS(A)$) et la qualité

d'ajustement aux données (R^2 cumulés). Compte tenu des critères précédents, la décision a été de choisir ici $A=1$. Il est clair que le fait de ne prendre qu'une seule composante interdit la représentation des individus dans des plans factoriels, représentation possible à l'aide de la fonction `plss.plot`. Les arguments d'entrée de cette fonction dont l'interface conversationnelle est en anglais sont décrits dans l'Annexe 2.

```
>plss.plot(essai0)
```

```
General information on the data? (y/n) 1: n
```

```
The B-spline transformations of the 10 predictors are now being computed
1  2  3  4  5  6  7  8  9  10
```

Il est possible d'afficher les statistiques univariées élémentaires pour les prédicteurs et les réponses ainsi que d'autres informations sur la nature des splines utilisées. Ensuite, le programme calcule les fonctions splines de base sur une grille de `nbpoints=40` points dans l'étendue de chaque prédicteur afin de représenter les graphes des fonctions coordonnées. Pour une discrétisation plus fine choisir une valeur plus élevée du paramètre `nbpoints` de la fonction `plss.plot`.

```
Y variance = 1 , reconstituted variance = 0.7537 according to components
```

	Y var.	%	cumul. %
1	0.7537	75.37	75.37

```
R2 of the reponses on the t subspaces according to the dimension
```

	Char
1	0.754

```
additive influence of the predictors on the t components? (y/n) 1: y
```

```
The number of the component t ?,(<= 1 ) 1: 1
```

```
Range of the transformed predictors in descending order:
```

```
COND( 1) Mg( 6) Ca( 5) Somme(10) SO4( 8) HCO3( 9) SiO2( 2) K( 4) Na( 3)
```

```

0.6922 0.6331 0.6161    0.6002  0.5784   0.1696   0.1278 0.0176 0.0103
Cl( 7)
0.0101

```

```

Hit <return> to see next plot

```

L’objectif est maintenant d’interpréter les composantes principales en essayant de dégager les prédicteurs de plus grande influence. L’influence d’une variable n’est en général pas linéaire sauf dans ce cas très particulier où PLSS se confond avec PLS. Le logiciel classe les variables explicatives selon l’étendue des données transformées. Cependant, ce classement est à considérer avec précaution car le paramètre de dispersion considéré est directement dépendant des valeurs extrêmes. Lorsque le nombre de prédicteurs n’est pas trop grand (≤ 30), un graphique, Figure 10, présente le diagramme en bâtons des étendues des transformations des échantillons.

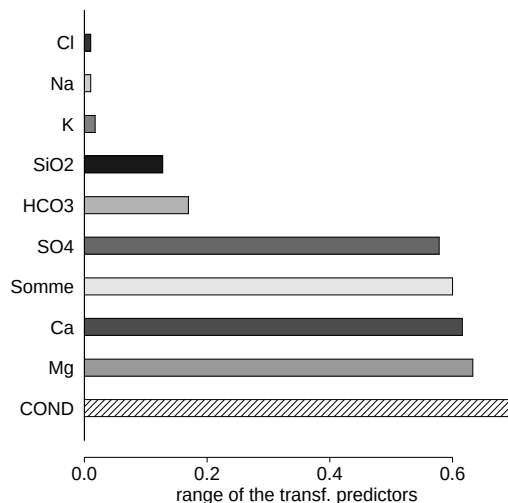


Figure 10: Influence d’un prédicteur sur t^1 selon l’étendue de sa transformation.

La fonction `plss.plot` propose d’examiner les graphes des fonctions coordonnées, dans ce cas linéaires, afin d’étudier leurs rôles dans la construction de t^1 .

```

Coordinate function plots of the ordered influential predictors? (y/n) 1: y
How many predictors?,(<= 10 ) 1: 10
How many plots on a line ?,(<= 10 ) 1: 4
How many line(s) ?1: 3
Do you need curves, data or both? (c,d,b) 1: b
Another t component ? (y/n) 1:n

```

Les 10 fonctions coordonnées sont présentées dans une matrice de graphiques sur 3 lignes $\times$ 4 colonnes. L'importance décroissante des variables se lit de la gauche vers la droite et de haut en bas. L'axe horizontal permet de lire les valeurs du prédicteur centré-réduit et l'axe vertical les valeurs centrées des variables transformées. Les 5 variables prépondérantes dans le modèle linéaire de t^1 doivent leur classement à quelques individus extrêmes tels que M , P et K . Ces individus sont ceux qui contribuent le plus à la variance de la première composante.

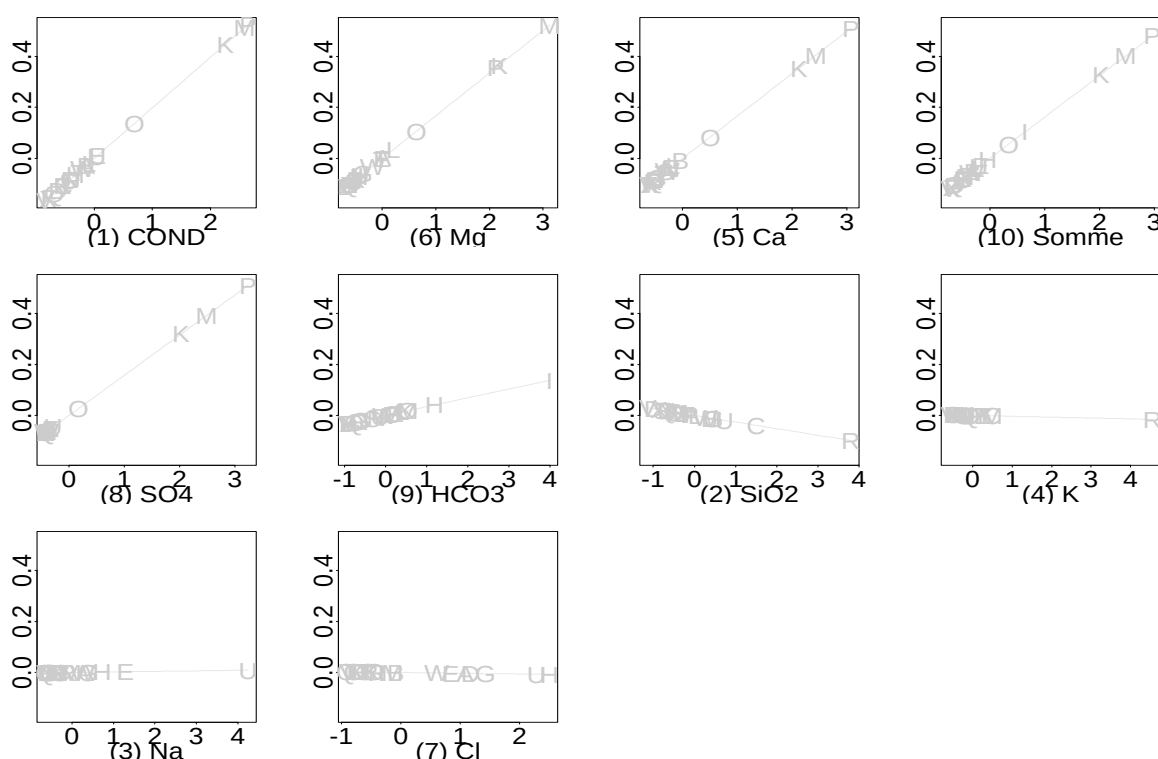


Figure 11: Graphe des 10 fonctions coordonnées de t^1 classées par influence décroissante de gauche à droite et de haut en bas.

Le type de graphiques de la Figure 11 fournit l'interprétation des composantes en fonction des variables explicatives. La Figure 12 présente la régression linéaire de u^1 avec t^1 . Dans le cas de la régression PLS uni-réponse, u^1 s'identifie avec la réponse et donc ici avec **Chargé**. Il est clair sur la Figure 12 que ce sont les jus d'orange M , K et P qui ont le plus d'influence sur le modèle. Dans le cadre multi-réponses ce type de graphique met en évidence la qualité globale des régressions partielles des Y^j sur t^1 , car u^1 est une combinaison linéaire des réponses.

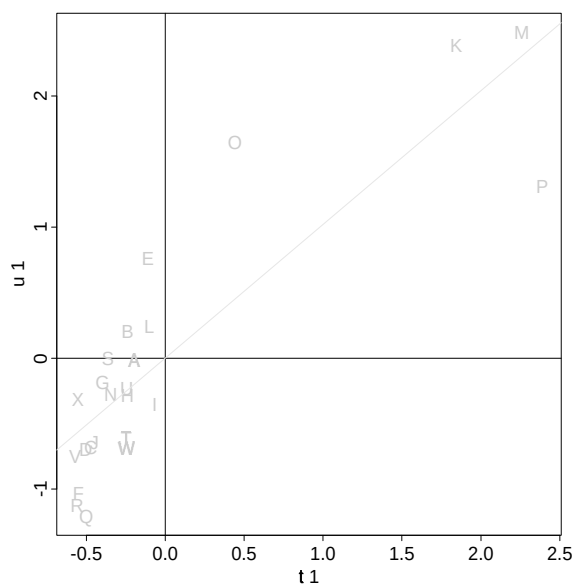


Figure 12: Représentation des données et de la droite de régression entre u^1 et t^1 .

Nous passons ensuite à l'examen du modèle PLSS de la variable **Chargé** en les prédicteurs, c'est à dire, ici, à la représentation des fonctions coordonnées linéaires.

additive influence of the predictors on a response variable? (y/n) 1: y

Influence of the predictors on the response: Char

Range of the transformed predictors in descending order:

```
COND( 1)  Mg( 6)  Ca( 5) Somme(10) SO4( 8) HCO3( 9) SiO2( 2)   K( 4)  Na( 3)
0.70681 0.64652 0.62917   0.61293 0.59062  0.17321  0.13047 0.01795 0.01053
Cl( 7)
0.01026
```

Coordinate function plots of the ordered influential predictors? (y/n)1: y

How many predictors?,(<= 10)1: 10

How many plots on a line ?,(<= 10)1: 4

How many line(s) ?1: 3

Do you need curves, data or both? (c,d,b)1: b

L'analyse par la Figure 13, du modèle linéaire de **Chargé** à 1 dimension, est reportée à plus tard pour être effectuée en comparaison avec celle du modèle spline de degré 2 à 2 dimensions de la Figure 15.

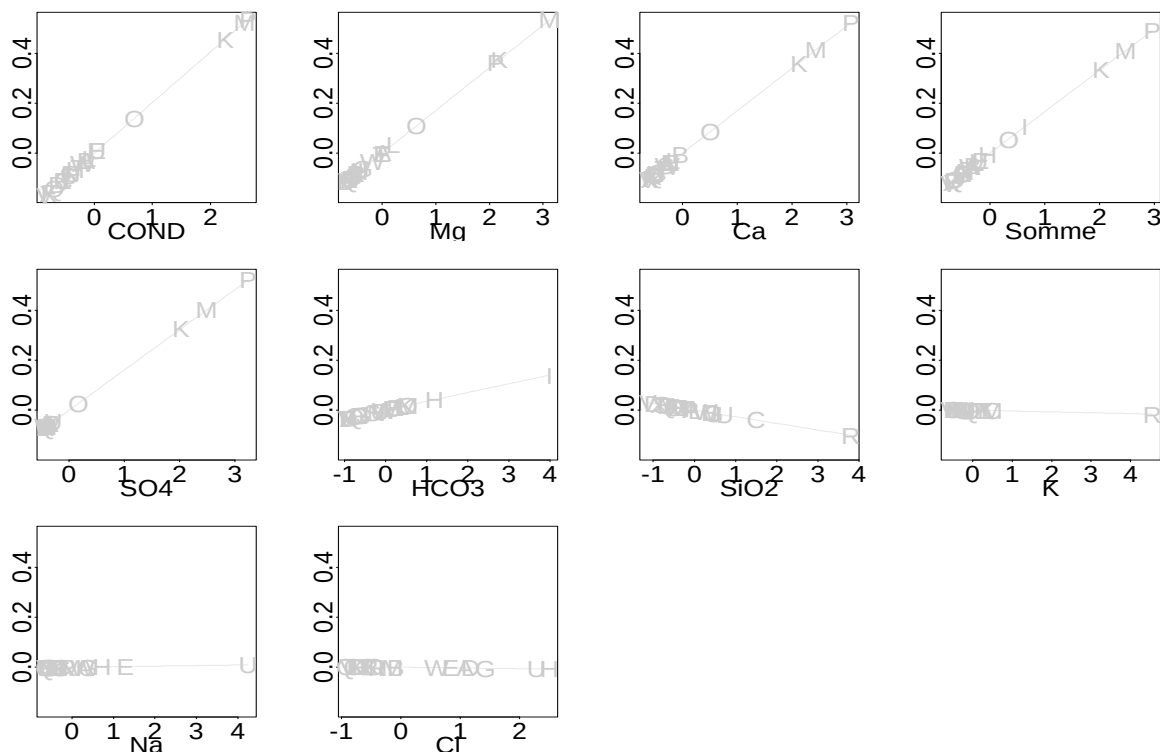


Figure 13: Modèle linéaire à 1 composante de la variable **Chargé** en les prédicteurs.

Remarquons pour en terminer avec cette vision spline additive du modèle linéaire que l'option `knots=0` jointe à `degree=3` dans les arguments de la fonction `plss` ainsi que dans `plssc`, permettent par exemple une modélisation PLS polynomiale de degré 3. Il est possible d'individualiser le type de transformation pour chaque prédicteur en affectant un vecteur de p valeurs à chacun de ces paramètres.

4.1.2 Augmentation du degré et du nombre de noeuds

PLS linéaire et l'examen des graphiques bivariés entre les réponses et les prédicteurs comme ceux des Figures 5 et 6, incitent à préférer les modèles polynomiaux par morceaux. Dans la stratégie du choix du degré et des noeuds, quelques règles heuristiques permettent de guider l'utilisateur. Il semble que le degré n'est pas le super-paramètre de plus grande influence, car l'augmentation du nombre de noeuds assure la souplesse d'ajustement qu'un degré trop faible risquerait de pénaliser. En effet, l'heuristique qui guide l'adjonction d'un noeud est la suivante :

ajouter un noeud augmente la liberté d'ajustement local au voisinage de la position choisie.

Le degré d'une spline n'excèdera pas 3 dans la plupart des cas. L'objectif de la détermination du nombre et de la position des noeuds est d'essayer d'éviter un sur-ajustement aux données tout en capturant des non-linéarités interprétables. Le contrôle de la nature des splines s'effectue comme pour la fonction `lss`. Pour localiser les noeuds, la valeur du paramètre `knots` grâce à l'option par défaut `equiknots=F`, place `knots` noeuds aux quantiles des variables, la valeur `equiknots=T` fournit `knots` noeuds distincts également espacés. Il est possible de spécifier la position des noeuds sur chaque variable grâce à l'option `listknots=list(c(...),...,c(...))`, où `c(...)` est un vecteur localisant les noeuds pour la variable concernée. Lorsque `listknots` est absent ce sont les options `knots` et `equiknots` qui sont actives.

L'essai suivant met en oeuvre la modélisation de **chargé** utilisant des splines de degré 2 avec des noeuds localisés afin de fournir de la liberté d'ajustement au voisinage des données extrêmes de certaines variables. L'examen de la Figure 5 montrant les graphes bivariés entre **Chargé** et les prédicteurs, a été un guide utile pour cet objectif. D'autre part, le principe d'économie de la dimension d'espace spline a essayé de nous guider.

```
>localnoeuds<-list(c(400,1600),c(10,20,40),c(10,40),c(2.5,5),c(160,400),
>+ c(40,110),c(4,11,30),c(400,1700),c(100,300,500),c(600,2600))
```

Après avoir décidé du type de splines adapté à chaque prédicteur, l'étape suivante consiste à déterminer le nombre de composantes par validation croisée. Une exploration de $A=7$ composantes possibles avec une proportion de 5% d'individus (= 1 individu) enlevés/prédits, est obtenue par les instructions suivantes.

```
>essai1cv<-plssc(juicX,juicY[,1,drop=F],degree=2,listknots=localnoeuds,
>+ A=7,prop=0.05)
>plscv.plot(essai1cv)
```

La dimension optimale du modèle PLSS spline est $A_{opt} = 2$ comme le montre la Figure 14 pour un $PRESS(A_{opt}) = 0.15$. Ces résultats sont à mettre en parallèle avec ceux de la Figure 9 obtenus par le modèle linéaire : meilleur $PRESS$ optimal associé à une convexité du graphe $PRESS(A)$ confirmant la qualité de la validation croisée.

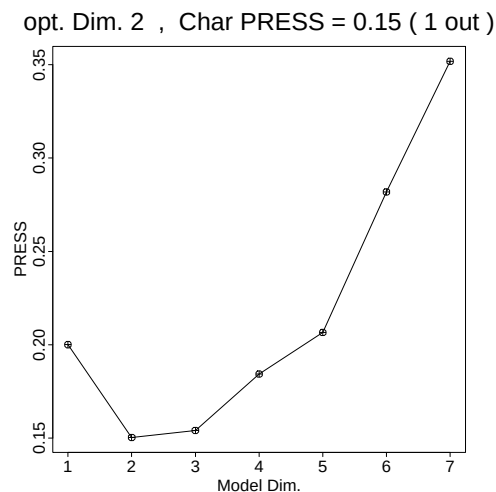


Figure 14: Validation croisée du modèle PLSS de Chargé.

Il reste à s'assurer de la qualité de l'ajustement du modèle à deux dimensions puis à exhiber le modèle spline additif obtenu pour les deux composantes t^1 , t^2 et pour la réponse chargé.

```
>essai1<-pls(juicX,juicY[,1,drop=F],degree=2,listknots=localnoeuds)
>plss.plot(essai1)
```

PLSS : PLS LINEAIRE SUR LE CODAGE SPLINE DES PREDICTEURS

- PLS lineaire -

Variance totale de Y = 1

Variance totale de B = 2.994889

Dimension 1

cov(t,u)=0.836444 r(t,u)=0.924 stdev(t)=0.9050577 stdev(u)= 1

Char % VAR Y

R2 part. 0.854 85.412

% VAR B expliquée par la comp. = 28.4809776241923

Dimension 2

cov(t,u)=0.1700213 r(t,u)=0.654 stdev(t)=0.6803583 stdev(u)=0.3819362

Char % VAR Y

R2 part. 0.062 6.245

R2 cum. 0.917 91.657

% VAR B expliquée par la comp. = 18.8182600637116

% VAR B expliquée = 47.2992376879039

nombre d'axes supplémentaires ?1: 0

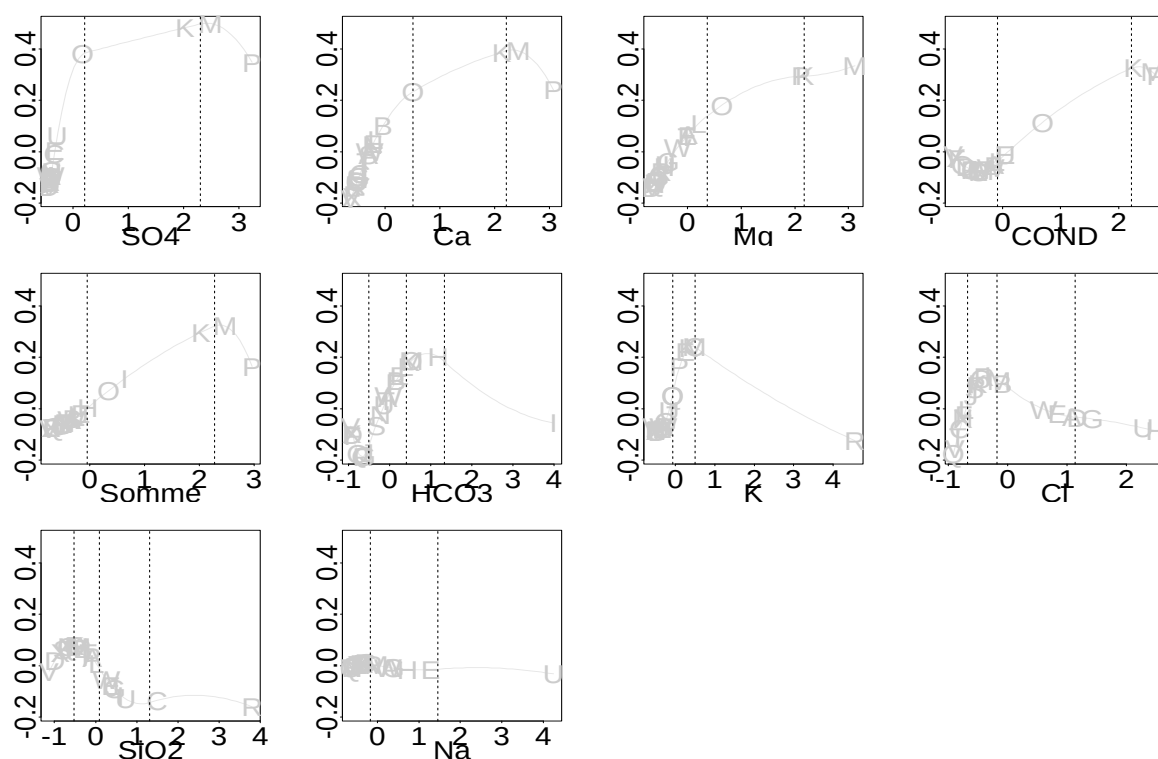


Figure 15: Modèle spline additif de **Chargé** avec 2 composantes. Les splines sont de degré 2 avec des noeuds indiqués par des verticales en pointillés.

La Figure 15 représente les fonctions coordonnées du modèle additif de la variable **Chargé**. Ce modèle s'ajuste bien aux données ($R^2 = 0.917$) et l'on peut remarquer que l'allure des fonctions coordonnées est en accord avec les graphes bivariés de la Figure 5 : PLSS rend bien compte de l'existence de liaisons non linéaires. L'interprétation de ce modèle apporte des nuances aux conclusions que le modèle linéaire de la Figure 13 propose. Comme dans le modèle linéaire, le caractère **Chargé** d'une boisson est principalement caractérisé par les variables SO_4 , Ca , Mg , $Cond$ et $Somme$ mais l'influence de ces variables ne peut plus être considérée comme linéaire sauf peut-être pour Mg . Les boissons ayant le plus de **Chargé**

sont M , K , P et O . D'autre part, les autres variables que les 5 déjà citées n'avaient quasiment aucune influence dans le modèle linéaire, la raison principale étant la présence de jus d'orange atypiques pour ces variables. Le fait que les bases de splines soient à support local rend plus robuste le modèle spline additif face au problème des données aberrantes. Par exemple, les données I et R aberrantes respectivement pour les variables HCO_3 et K ne perturbent pas trop le rôle des autres jus d'orange dans la construction des fonctions coordonnées de ces variables (la position des noeuds est cruciale).

4.2 Régression multi-réponses de l'Intensité et de l'Agressivité

Les quatre variables Int_1 , Int_2 , Agr_1 et Agr_2 étant très fortement corrélées positivement, il est préférable de les modéliser par PLSS simultanément plutôt que séparément. L'accent sera mis sur les potentialités des fonctions `plss` et autres, non présentées dans la section précédente. En particulier, nous ne présenterons pas les différents modèles des 4 réponses mais seulement celui de la variable Int_1 . A partir d'un jeu de données test, une étape de prédiction-validation "externe" est présentée dans la section suivante. Enfin, l'exploration des données par les plans factoriels PLSS fait l'objet de la dernière partie.

Il est très utile de visualiser par `bivarplot` les graphes bivariés de chaque réponse avec les prédicteurs afin de se donner des idées sur les types de splines à utiliser. Le degré choisi a été 2. Bien que déterminer le nombre et la position des noeuds soit plus difficile encore dans le cadre multi-réponses, le fait que les variables à expliquer soient ici très corrélées, rend la tâche un peu moins ardue. En présence de données extrêmes comme c'est le cas, il est préférable de localiser les noeuds d'une façon adaptée à chaque prédicteur plutôt que également espacés ou bien aux quantiles.

```
>localnoeuds1<-list(c(400,1600),20,30,c(2,5),c(100,400),40,20,
>+ c(100,1500),c(200,300,400),c(500,2500))
```

4.2.1 Prédiction-validation sur un jeu de données test

Une fonction spline étant nulle en dehors de l'intervalle $[a, b]$ sur lequel elle a été construite, il est nécessaire de contrôler si les observations à prédire sont dans l'étendue des valeurs des variables explicatives qui ont servi à bâtir le modèle. Le jeu de données

test utilisé ici est constitué par les deux seuls jus d'orange K et L . Il est clair qu'une validation "externe" à partir d'un deuxième jeu de données, doit comporter un nombre suffisant d'observations. Ici, l'objectif est seulement d'illustrer les possibilités de prédiction-validation externe par la fonction `plss`. Aussi, une étape de validation croisée est préalablement effectuée sur les données privées des boissons K et L .

```
>essai2cv<-plsscvc(juicX[-c(11,12),],juicY[-c(11,12),2:5],degree=2,
>+ listknots=localnoeuds1,A=7,prop=0.05)
>plscv.plot(essai2cv)
```

Le résultat de la validation croisée fournit 3 composantes pour un PRESS total (somme des PRESS des 4 réponses) ayant la valeur 1.72. Il est à noter que les variables d'agressivité sont mieux prédites par ce modèle que les variables d'intensité.

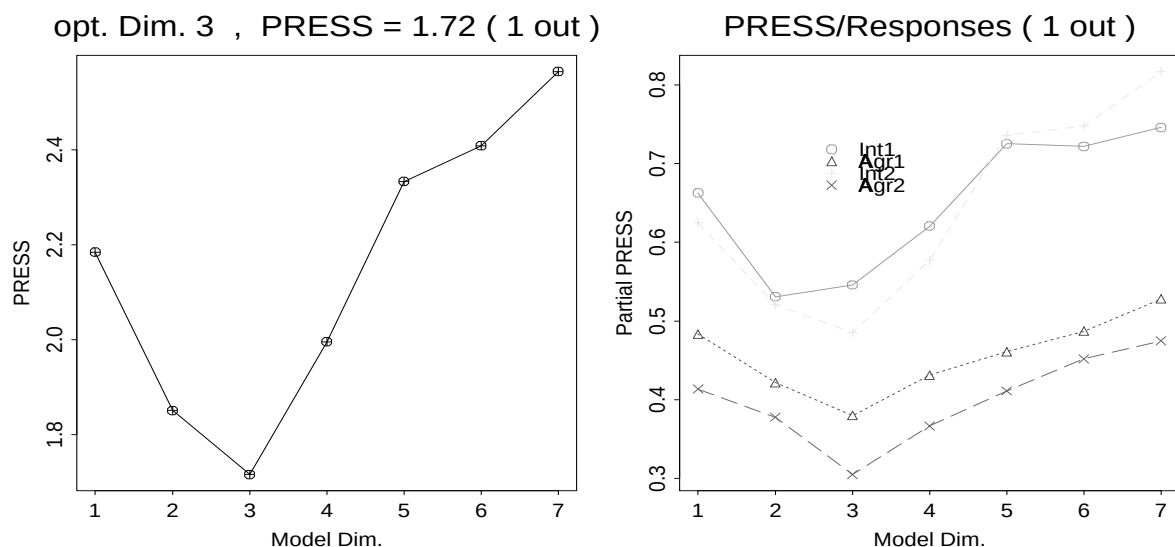


Figure 16: PRESS total et PRESS de chaque réponse du modèle PLSS.

Pour une étape de validation-prédiction, les arguments d'entrée `Xtest`, `Ytest` de `plss` sont affectés par les matrices des échantillons du deuxième jeu de données. Lorsque `Ytest` est absent, le modèle est utilisé pour faire de la prédiction à partir de `Xtest`.

```
>plss(juicX[-c(11,12),],juicY[-c(11,12),2:5],Xtest=juicX[c(11,12),],
>+ Ytest=juicY[c(11,12),2:5],degree=2,listknots=localnoeuds1,A=3)
```

Validation par l'échantillon test suivant la dimension ? (o/n)1: o

Moyenne des erreurs suivant la dimension

	1	2	3
MSE	2.393	1.738	2.103

Choisissez la dimension du modele (<= 3)1: 2

Reponses estimees et erreurs de prediction Yerr = Ytest - Yest

	est.Int1	err.Int1	est.Agr1	err.Agr1	est.Int2	err.Int2	est.Agr2	err.Agr2
K	14.1114	2.2086	9.9186	1.4814	15.1695	2.7705	10.4867	0.4333
L	7.3480	1.4420	4.0554	1.7046	8.6536	1.5164	4.0652	1.724

Pour une dimension fixée du modèle, ici ≤ 3 , l'erreur moyenne d'une réponse est la racine carrée de la moyenne des carrés des erreurs sur chaque individu. L'erreur globale sur les q réponses, MSE, est la moyenne des erreurs précédentes. Compte tenu de la petite taille de l'échantillon test, La dimension 2 choisie ne peut être retenue comme dimension optimale.

4.2.2 Exploration des données et modélisation

Les composantes principales PLSS retenues permettent une visualisation des individus dans des plans factoriels. L'explication et l'interprétation des composantes par les prédicteurs passe par l'examen du modèle spline additif

$$t^k = \sum_{i=1}^p \varphi_i^k(x^i), \quad k = 1, \dots, A,$$

dont la relation (11) est la version matricielle sur l'échantillon des données. Cela signifie que, excepté le cas par défaut où le modèle PLSS est linéaire, les composantes s'expriment comme une somme de fonctions non linéaires des variables explicatives. Les fonctions coordonnées $\varphi_i^k(\cdot)$ de plus grande influence permettent de détecter les individus qui contribuent le plus à la variance de la composante t^k . Comme pour le modèle PLSS des réponses, l'influence de la fonction coordonnée $\varphi_i^k(\cdot)$ est évaluée par l'étendue des valeurs qu'elle prend sur l'échantillon de la variable explicative x^i .

Pour interpréter les composantes par les réponses, PLSS étant la régression PLS linéaire sur les codages splines des prédicteurs, la relation (7) justifie l'utilisation des corrélations linéaires des réponses avec les composantes pour interpréter ces dernières.


```
>essai2<-plss(juicX,juicY[,2:5],degree=2,listknots=localnoeuds1,
>+ A=3,impres=F)
>plss.plot(essai2)
```

L'option `impres=F` supprime le mode conversationnel. La validation croisée sur le jeu de données complet, non présentée ici, confirme le résultat, $A = 3$ composantes, obtenu sans les deux boissons K et L .

Additive influence of the predictors on the t components? (y/n)1: y

The number of the t component ?,(<= 3)1: 1

Range of the transformed predictors in descending order:

```
COND( 1) Somme(10) HCO3( 9) Ca( 5) SiO2( 2) Mg( 6)  K( 4) SO4( 8) Na( 3)
0.5361      0.533      0.492 0.4468      0.348 0.3013 0.2816      0.267 0.1762
Cl( 7)
0.1351
```

Coordinate function plots of the ordered influential predictors? (y/n)1: y

How many predictors?,(<= 10)1: 4

How many plots on a line ?,(<= 4)1: 4

How many line(s) ?1: 1

Do you need curves, data or both? (c,d,b)1: b

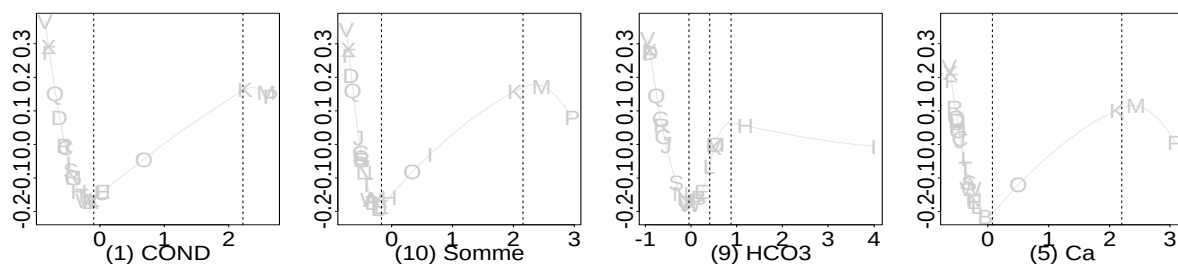


Figure 17: Les 4 premières fonctions coordonnées de t^1 classées par ordre décroissant.

La Figure 17 permet de mettre en évidence les boissons qui contribuent le plus à la variance de t^1 . Les trois individus V , X et F qui ont les plus faibles valeurs sur les 4 prédicteurs $COND$, $Somme$, HCO_3 et Ca , sont ceux qui ont la plus grande influence sur la première composante.

De la même façon, on peut sélectionner 8 prédicteurs influents sur t^2 .

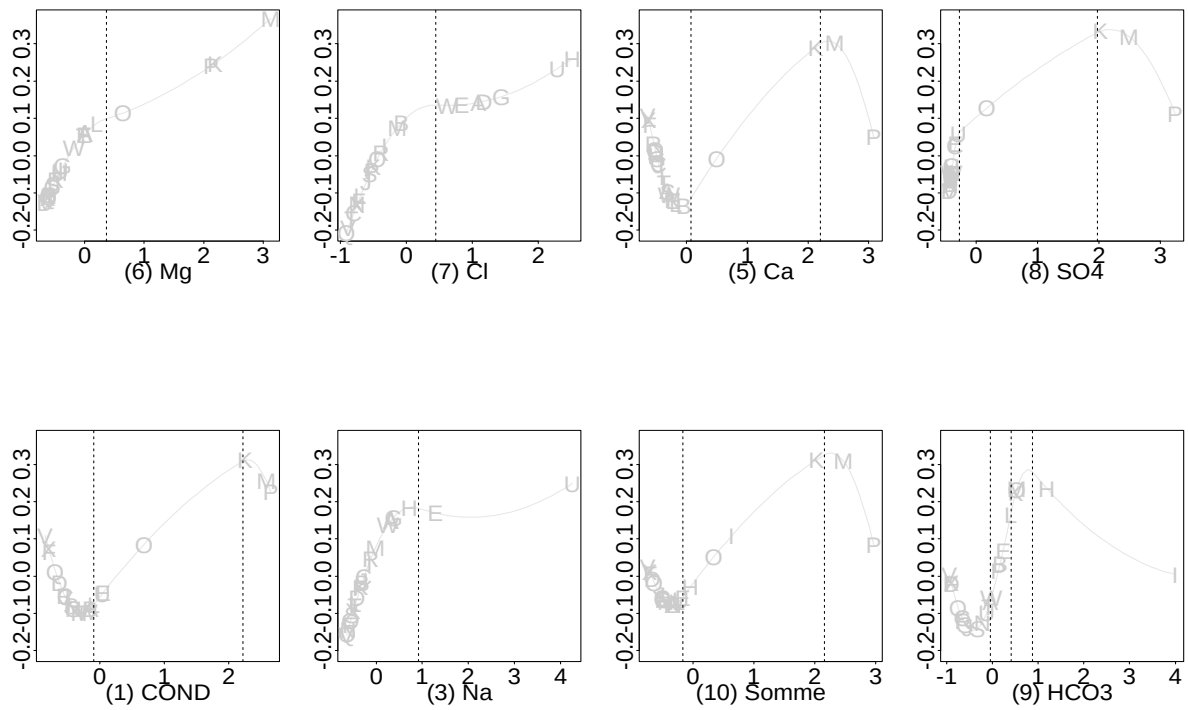


Figure 18: Huit premières fonctions coordonnées de t^2 classées par ordre décroissant.

C'est la deuxième composante principale qui met en évidence les boissons M , K et à un degré moindre P comme individus influents. Les Figures 17 et 18 permettent d'interpréter t^1 et t^2 en les prédicteurs. Les individus peuvent alors être visualisés, Figure 19, dans le plan factoriel (t^1, t^2).

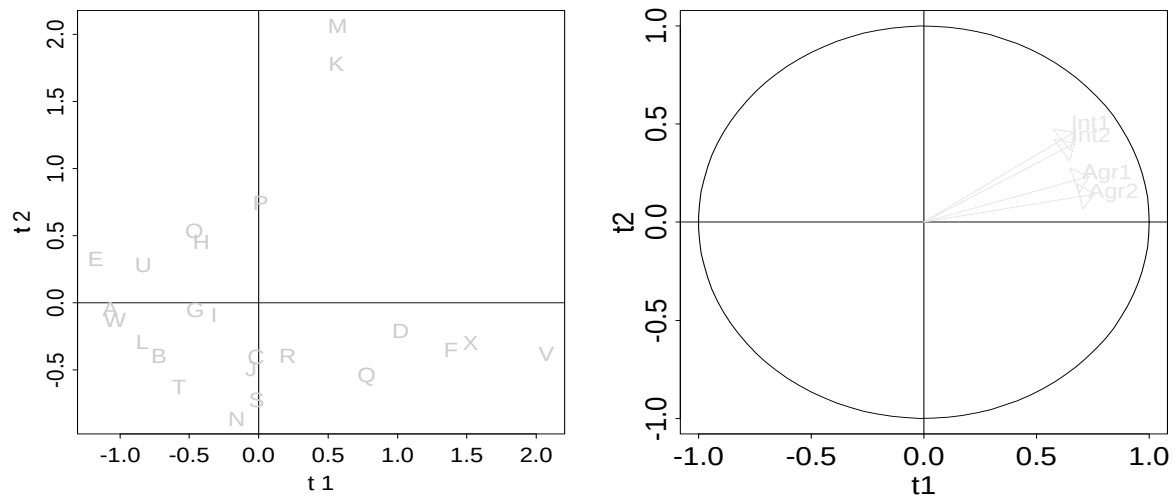


Figure 19: Représentation des individus dans le premier plan factoriel et cercle des corrélations pour les réponses.

Les corrélations entre les réponses et les composantes montrent que le premier axe est fortement positivement corrélé avec les variables d'Intensité et d'Agressivité.

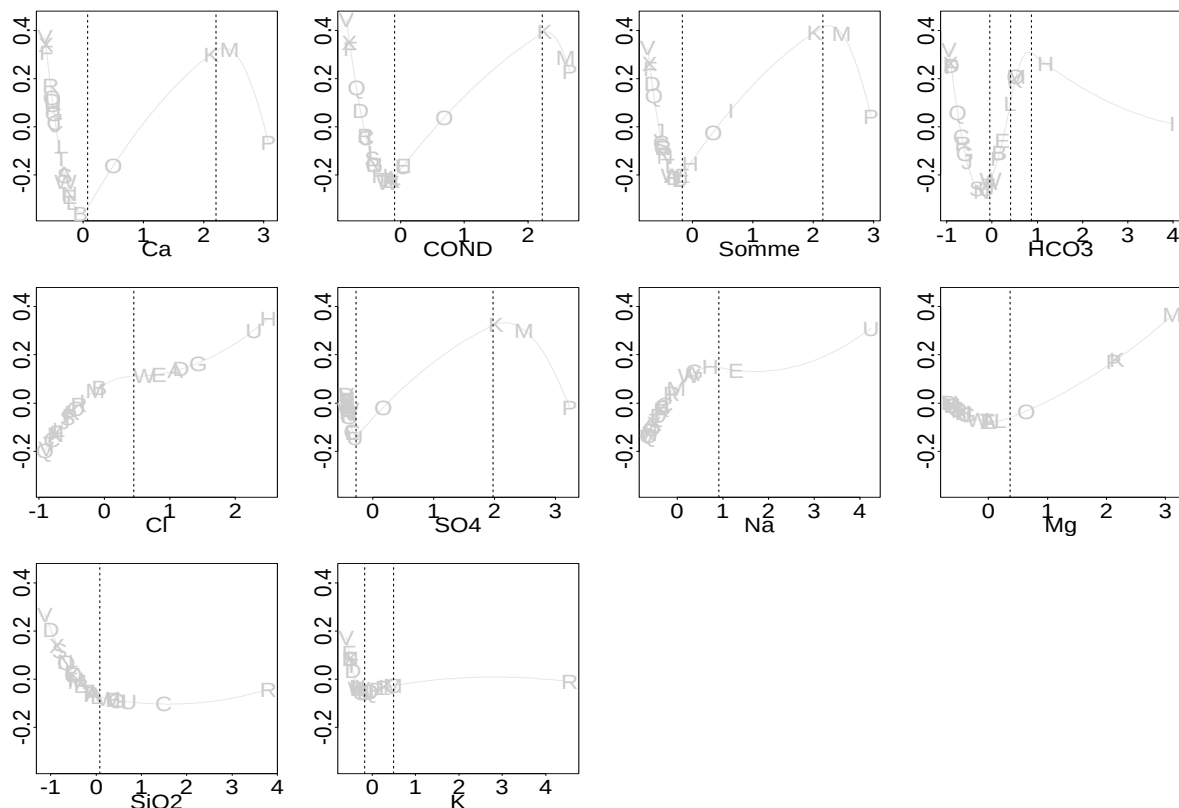


Figure 20: Modèle PLSS de la variable Int_1 à 3 composantes. Les splines sont de degré 2 et les noeuds sont localisés par des verticales en pointillés.

Une analyse fine des modèles des variables d'intensité et d'agressivité nécessiterait plus de développements et d'explications sur la nature des variables réponses et sur les données en général. Cependant, le modèle de la variable Int_1 est très fortement non linéaire en les trois prédictors de plus grande influence Ca , $COND$ et $Somme$. Des boissons comme V , X et F ayant des faibles valeurs sur ces prédictors présentent une forte intensité Int_1 ainsi que d'autres comme K et M qui prennent de fortes valeurs sur ces variables, la boisson P présentant cependant une intensité moyenne. Par contre, les boissons qui prennent des valeurs moyennes sur ces variables, voisinage de 0 sur l'axe des abscisses, sont celles dont la valeur Int_1 est faible.

5 Quelques dernières remarques

Les données des jus d'orange qui ont servi à montrer les potentialités des différentes fonctions S-Plus du modèle PLSS, présentent un contexte difficile pour la régression où la rareté des observations s'ajoute à une forte corrélation entre certaines variables explicatives.

La méthode PLSS a permis d'une part de vérifier que le modèle PLS linéaire (option par défaut de `plss`) n'était pas bien adapté et d'autre part de proposer un modèle additif polynomial par morceaux présentant une meilleure qualité de prédiction. L'une des raisons de ce succès réside certainement dans le fait que certaines variables explicatives présentent des valeurs extrêmes à forte influence sur les composantes PLS linéaires. L'utilisation des bases de B -splines à support local, permet par un choix adapté des noeuds, de réduire l'influence de telles données sur les composantes ainsi que sur le modèle.

L'utilisateur de PLSS dispose d'un large éventail de modèles PLS allant des modèles polynomiaux classiques aux modèles polynomiaux par morceaux. Il n'existe pas dans l'état actuel du développement de la méthode, de gestion automatique des superparamètres des fonctions splines. Les choix du degré des polynômes, du nombre et de la position des noeuds, sont du ressort de l'utilisateur. En particulier, la localisation de noeuds est souvent prépondérante pour faire ressortir telle variable explicative comme plus ou moins influente dans le modèle additif. Les options par défaut de noeuds équidistants ou de noeuds aux quantiles permettent une première exploration des modèles.

Enfin, PLSS hérite de PLS linéaire des propriétés qui font de cette méthode à la fois un outil de modélisation et d'exploration non linéaire des données. L'interprétation des plans factoriels passe par l'examen des fonctions coordonnées qui interviennent additivement dans les composantes PLSS. L'utilisateur devra se familiariser avec cette approche qui n'est pas trop éloignée, dans la pratique, des méthodes usuelles de l'Analyse des Données.

6 ANNEXE

6.1 Annexe 1 : Données des jus d'orange

Prédicteurs : variables minéralogiques

	COND	SiO2	Na	K	Ca	Mg	Cl	S04	HC03	Somme
A	376.0	17.29	19.20	1.43	44.80	26.00	29.20	18.20	199.12	355.24
B	369.6	13.39	7.52	1.43	83.20	2.34	12.40	6.50	246.96	373.74
C	168.8	42.90	8.32	4.94	24.00	4.55	2.08	54.60	63.44	204.83
D	121.6	2.34	5.60	0.78	17.20	0.00	30.40	0.00	13.68	70.00
E	472.8	24.70	36.80	3.90	58.40	25.48	25.60	60.45	261.60	496.93
F	40.8	9.75	2.56	0.39	4.00	2.21	3.12	5.20	14.64	41.87
G	234.4	26.00	20.16	1.69	21.12	11.57	34.16	26.00	80.00	220.70
H	280.0	11.57	26.40	1.30	19.20	8.97	49.60	24.70	468.48	610.22
I	200.0	28.60	10.40	1.30	32.80	10.01	8.80	18.20	1079.44	1189.55
J	168.0	25.35	7.76	1.95	24.88	5.98	4.80	13.00	89.76	173.48
K	1618.4	10.40	6.96	4.29	388.80	110.50	6.40	1534.00	321.12	2382.47
L	398.4	18.85	4.80	1.30	64.00	33.54	3.52	16.25	297.68	439.94
M	1788.0	11.96	12.80	4.94	432.00	146.90	11.52	1820.00	326.00	2766.12
N	234.4	7.54	1.92	0.52	56.80	2.08	2.88	20.80	161.04	253.58
O	804.0	10.40	2.96	2.47	160.00	50.70	7.28	377.00	325.04	935.85
P	1823.2	16.90	6.80	3.25	522.40	107.90	6.00	2288.00	239.12	3190.37
Q	88.8	7.80	1.52	2.47	16.16	1.95	0.56	16.90	44.88	92.24
R	165.6	80.60	11.04	23.27	12.88	6.76	8.00	7.80	70.24	220.59
S	216.0	5.59	3.60	0.52	45.60	5.07	5.60	10.40	139.60	215.98
T	301.6	15.60	4.56	0.65	37.60	24.70	1.60	15.60	191.28	291.59
U	468.0	30.42	91.20	2.08	18.24	9.88	46.40	88.40	182.48	469.10
V	8.8	0.26	1.44	0.13	0.16	0.26	0.88	0.00	4.88	8.01
W	341.6	22.88	17.92	1.43	48.80	18.59	22.40	17.55	204.96	354.53
X	34.4	4.68	4.00	0.52	1.92	1.56	2.80	5.20	12.72	33.40

Réponses : variables sensorielles

	Char	Int1	Agr1	Int2	Agr2
A	9.81	11.73	8.46	11.97	7.86
B	10.68	9.84	6.96	12.63	7.83
C	7.11	11.01	6.09	11.97	6.72
D	7.05	13.05	10.56	15.21	11.67
E	12.96	8.88	4.65	9.45	3.42
F	5.70	14.25	13.83	15.99	14.52
G	9.12	9.24	4.98	11.82	6.39
H	8.73	12.24	9.15	13.65	10.14
I	8.43	10.59	8.13	11.73	9.21
J	7.26	10.02	6.09	11.52	6.21
K	19.53	16.32	11.40	17.94	10.92
L	10.86	8.79	5.76	10.17	5.79
M	19.92	17.97	12.15	17.88	12.69
N	8.76	9.45	5.82	9.00	5.61
O	16.53	12.66	6.96	13.38	6.66
P	15.18	11.73	7.20	12.18	6.87
Q	4.98	12.03	9.00	13.08	8.79
R	5.31	11.70	8.04	13.35	9.75
S	9.87	9.42	6.00	10.44	6.18
T	7.38	10.11	5.82	10.62	6.48
U	8.94	12.63	7.44	15.45	8.01
V	6.84	17.67	13.95	19.20	17.07
W	7.08	11.22	8.10	11.97	8.04
X	8.58	17.61	15.42	19.11	16.53

6.2 Annexe 2 : Description des fonctions S-Plus implémentées

Cette annexe présente les fonctions S-Plus non natives implémentées par l'auteur et correspondant à la version 9.6 du logiciel. Pour obtenir une aide en ligne des fonctions natives, utiliser les fonctions `help()` ou `help.start()`.

↪ bivarplot

Cette fonction fournit les graphes bivariés entre une variable réponse (une colonne de Y) et les variables explicatives (les colonnes de X). Les observations sont représentées ainsi qu'un lissage non linéaire par une méthode non-paramétrique. Voir aussi `pairs()`.

<i>entrées</i>	<i>description</i>	<i>défaut</i>
X, Y	matrices des var. explicatives et des réponses.	
matrow, matcol	entiers; nb de lignes et de colonnes de la matrice des plots.	1, 1
smooth	caractères; le nom du lisseur utilisé parmi <i>lowess</i> , <i>ksmooth</i> , <i>supsmu</i> , <i>smooth.spline</i> .	"lowess"
typedata	booléen; si T, les plots donnent le nom des observations, si F, un indicateur paramétré par pchpar.	T
ptypar, cexpar, pchpar	valeurs des param. S-Plus pty, cex et pch pour les plots.	"s" ; 0.7; "*"
titlepar	booléen; si T, met un titre sur les plots multiples.	T
colpar	entier; numéro de la couleur pour les plots.	1

↪ lss

Régression "Least-Squares Splines" dans le cadre univarié ou dans le cadre multivarié additif (fonctions coordonnées). Les individus peuvent être individuellement pondérés.

<i>entrées</i>	<i>description</i>	<i>défaut</i>
X,Y	matrices des var. explicatives et des réponses.	
degree, knots	vecteurs de ncol(X) entiers: degrés, nombres de noeuds; si une seule valeur: toutes les variables prennent cette valeur; degree=1, knots=0 pour la régression linéaire.	1; 0
equiknots	vecteur booléen de ncol(X) valeurs; si une seule valeur: répétée pour toutes les variables; T, noeuds équi-espacés, F, noeuds aux knots-quantiles des prédicteurs.	F
listknots	liste de ncol(X) vecteurs des localisations des noeuds; si manquant, knots et equiknots sont actifs.	manquant
D	vecteur des poids statistiques, $D = 1$ poids équi-répartis.	1
centerX, centerY	booléens; si T, centrage des variables.	F; F
standX, standY	booléens; si T, standardisation des variables.	F; F
invpar	booléen; si T, $(B'diag(D)B)^{-1}$, si F, Moore-Penrose.	T
impres	booléen; si F, supprime le mode conversationnel.	T
nbpoints	entier; nb de points de discrétisation des fonct. splines.	50
ptypar, cexpar, pchpar	idem dans bivarplot.	"s" ; 0.7; "*"
colpar, titlepar	idem dans bivarplot.	1; T
ncolpar	entier donnant le nombre de colonne de la légende.	1

Sorties :

\$ffit : moyenne des carrés des résidus dans le cas uniréponse, somme des moyennes des carrés des résidus dans le cas multi-réponses.

\$beta : matrice des coefficients splines de l'estimateur LSS.

\$Bcod : super-matrice de codage.

↪ plssc

Pour un choix de splines fixé, cette fonction calcule les valeurs du PRESS selon le nombre de composantes, pour la validation croisée du modèle PLSS, voir Section 3.3.3. Les individus peuvent être individuellement pondérés.

<i>entrées</i>	<i>description</i>	<i>défaut</i>
X,Y	matrices des var. explicatives et des réponses.	
A	entier; nombre maximal de composantes envisagé.	2
degree, knots	vecteurs de ncol(X) entiers: degrés, nombres de noeuds; si une seule valeur: toutes les variables prennent cette valeur; degree=1, knots=0 pour PLS linéaire.	1; 0
equiknots	vecteur booléen de ncol(X) valeurs; si une seule valeur: répétée pour toutes les variables; T, noeuds équi-espacés, F, noeuds aux knots-quantiles des prédicteurs.	F
listknots	liste de ncol(X) vecteurs des localisations des noeuds; si manquant, knots et equiknots sont actifs.	<i>manquant</i>
D	vecteur des poids statistiques; si $D = 1$ poids équi-répartis.	1
standX, standY	booléens; si T, standardisation des variables; F, centrage.	T; T
prop	réel; proportion des individus écartés/prédits.	0.1

Comme `plscv` qui effectue la validation croisée du modèle PLS linéaire, cette fonction n'a pas de sorties graphiques. Les sorties sont numériques et doivent être visualisées grâce à la fonction suivante.

↪ plscv.plot

Cette fonction donne le(s) graphique(s) du PRESS et le nombre optimal de composantes.

<i>entrées</i>	<i>description</i>	<i>défaut</i>
plscvresul	c'est l'objet S-Plus des sorties de <code>plscv</code> ou <code>plssc</code> .	
ptypar, cexpar, pchpar	idem dans <code>bivarplot</code> .	"s" ; 0.7; "*"
colpar, titlepar	idem dans <code>bivarplot</code> .	1;T

Dans le cas multi-réponses, le premier plot donne le PRESS total (somme des PRESS partiels pour chaque réponse). Le deuxième donne le PRESS partiel, moyenne des carrés des erreurs de prédiction, pour chaque réponse. Pour placer la légende, cliquer sur ce plot au point choisi comme coin supérieur gauche de la légende.

↪ **plss**

Régression “PLS Splines” dans le cadre uniréponse ou multi-réponses. Les individus peuvent être pondérés individuellement dans le calcul des covariances.

<i>entrées</i>	<i>description</i>	<i>défaut</i>
X,Y	matrices des var. explicatives et des réponses.	
Xtest, Ytest	matrices d'un échantillon test pour prédire ou valider.	<i>manquant</i>
A	entier; nombre de composantes ou dimension du modèle.	2
degree, knots	vecteurs de ncol(X) entiers: degrés, nombres de noeuds; si une seule valeur: toutes les variables prennent cette valeur; degree=1, knots=0 pour PLS linéaire.	1; 0
equiknots	vecteur booléen de ncol(X)valeurs; si une seule valeur: répétée pour toutes les variables; T, noeuds équi-espacés, F, noeuds aux knots-quantiles des prédicteurs.	F
listknots	liste de ncol(X) vecteurs des localisations des noeuds; si manquant, knots et equiknots sont actifs.	<i>manquant</i>
D	vecteur des poids statistiques; si $D = 1$ poids équi-répartis.	1
standX, standY	booléens; si T, standardisation des variables; F, centrage.	T, T
eps	réel positif; seuil pour le zéro des valeurs singulières.	1exp-08
impres	booléen; si F, supprime le mode conversationnel.	T

Principales sorties :

\$compX , **\$compY** : matrices des composantes t et u .

\$cX , **\$cY** : matrices des poids w et c .

\$BETALCR : liste des A matrices des coefficients splines de l'estimateur PLSS.

\$Bcod : super-matrice de codage.

↪ **plss.plot**

Cette fonction présente les sorties numériques et graphiques de **plss**.

<i>entrées</i>	<i>description</i>	<i>défaut</i>
resulpls	c'est l'objet S-Plus des sorties de plss.	
nbpoints	entier; nb de points de discrétisation des fonct. splines.	50
typedata	idem dans bivarplot.	T
ptypar, cexpar, pchpar	idem dans bivarplot.	"s" ; 0.7; "*"
colpar, titlepar	idem dans bivarplot.	1;T

6.3 Annexe 3 : Un peu de Mathématique sur les B-splines

L'espace vectoriel $S(m, \{\tau_{m+1}, \dots, \tau_{m+K}\}, [a, b])$ de dimension $r = m+K$, des fonctions splines polynomiales d'ordre $m = d+1$ ou de degré d , sur un intervalle $[a, b]$, est caractérisé par

- le nombre K de noeuds intérieurs à $[a, b]$,
- la position des noeuds intérieurs notés $\{\tau_{m+1}, \dots, \tau_{m+K}\}$

$$\tau_1 = \dots = \tau_m = a < \tau_{m+1} \leq \dots \leq \tau_{m+K} < b = \tau_{m+K+1} = \dots = \tau_{2m+K}.$$

On appelle multiplicité d'un noeud, le nombre de noeuds confondus en un même point. Cette multiplicité peut varier de 1, noeud simple, à m pour un noeud multiple d'ordre m . Une base de $S(m, \{\tau_{m+1}, \dots, \tau_{m+K}\}, [a, b])$ est constituée par les fonctions B -splines polynomiales normalisées et notée $\{B_j^m(\cdot)\}_{j=1, m+K}$. Ces fonctions sont notées plus simplement $B_j(\cdot)$ lorsqu'il n'y a pas d'ambiguïté, ce qui est le cas dans les sections précédentes.

Définition : Fonction puissance tronquée

La fonction de $\mathbb{R}$ dans $\mathbb{R}^+$, $x \longrightarrow (x - \tau)_+^k$ définie par $(x - \tau)^k$ si $x \geq \tau$, 0 si non, est appelée puissance tronquée de degré k (k entier non négatif). Ces fonctions peuvent être directement utilisées pour construire une base de l'espace des fonctions splines. Cependant, la base la plus utilisée est celle des B -splines.

Définition : Fonctions B -splines normalisées d'ordre m (degré $m - 1$)

Pour $j = 1, \dots, m + K$,

$$B_j^m(x) = (-1)^m (\tau_{j+m} - \tau_j) [\tau_j, \dots, \tau_{j+m}] (x - \tau)_+^{m-1}$$

où $[\tau_j, \dots, \tau_{j+m}] (x - \tau)_+^{m-1}$ est la différence divisée d'ordre m pour les valeurs $\tau_j, \dots, \tau_{j+m}$ de la fonction $\tau \longrightarrow (x - \tau)_+^{m-1}$.

Une spline, élément de $S(m, \{\tau_{m+1}, \dots, \tau_{m+K}\}, [a, b])$, s'écrit dans la base de ces fonctions

$$s(x) = \sum_{j=1}^{m+K} \beta_j B_j^m(x).$$

Propriété : Formule de récurrence (De Boor 1978)

Dans la pratique, on a seulement besoin de calculer la valeur d'une B -spline en un point. Il existe pour ce faire la formule de récurrence

$$\begin{aligned} B_j^1(x) &= 1 \text{ si } \tau_j \leq x \leq \tau_{j+1}, \quad 0 \text{ si non,} \\ \text{pour } k &= 2, \dots, m, \\ B_j^k(x) &= \frac{x - \tau_j}{\tau_{j+k-1} - \tau_j} B_j^{k-1}(x) + \frac{\tau_{j+k} - x}{\tau_{j+k} - \tau_{j+1}} B_{j+1}^{k-1}(x). \end{aligned}$$

Les principales propriétés de la base des B -splines sont présentées dans la Section 2.1.

7 BIBLIOGRAPHIE

- DE BOOR, C. (1978), *A practical guide to splines*. New York: Springer-Verlag.
- DURAND, J.F. (1993), Generalized principal component analysis with respect to instrumental variables via univariate spline transformations, *Computational Statistics & Data Analysis*, 16, 423-440.
- DURAND, J.F. (1997), *Additive Modeling of multivariate data by spline fonctions*, Rapport de Recherche n.97-04, Unité de Biométrie.
- DURAND, J.F. (1999), PLS and multivariate additive spline modeling, *Les Méthodes PLS, Symposium International PLS'99*, Tenenhaus, M. et Morineau, A. (Eds), CISIA-CERESTA, 1-20.
- DURAND, J.F., SABATIER, R. (1997), Additive Splines for PLS regression, *Journal of the American Statistical Association*, Vol. 92, 440, 1546-1554.
- DURAND, J.F., ROMAN, S. et VIVIEN M. (1998), *Guide de la régression PLS linéaire sous S-Plus*. Rapport de Recherche 98-06, Groupe de Biostatistique et d'Analyses des Systèmes, ENSA.M-INRA-UM II.
- HASTIE, T. and TIBSHIRANI, R. (1990), *Generalized additive models*, London: Chapman and Hall.
- MATHSOFT (1996), *S-Plus Version 3.4 for Unix Supplement*, Data Analysis Products Division, MathSoft, Seattle.
- TENENHAUS, M. (1998) *La régression PLS, théorie et pratique*, Paris : Technip.
- TENENHAUS, M. GAUCHI, J.P., et MÉNARDO, C. (1995), Régression PLS et applications, *Revue de Statistique Appliquée*, XLIII, 7-63.
- VENABLES, W.N. and RIPLEY, B.D. (1994), *Modern Applied Statistics with S-Plus*, New York: Springer-Verlag.
- WOLD, H. (1966), Estimation of principal component and related models by iterative least squares, *Multivariate Analysis*, ed.P.R.Krishnaiah, Newyork : Academic Press, 391-420.
- WOLD, H. (1975), Soft Modelling by latent variables; the nonlinear iterative partial least squares approach. In: *Perspectives in Probability and Statistics*, Gani, J. (Ed), (Papers in honour of M.S. Bartlett). London: Academic Press.

WOLD, S., MARTENS, H. and WOLD, H. (1983), The multivariate calibration problem in chemistry solved by the PLS method, *Proc. Conf. Matrix Pencils*. Ruhe, A. and Kagstrom, B. (Eds), Lecture notes in mathematics, Heidelberg : Springer Verlag, 286-293.