"Chemometrics & Intelligent Laboratory Systems, 58 (2001), 235-246. "

Local polynomial additive regression through PLS and Splines: PLSS

Jean-François Durand

Laboratoire de Probabilités et Statistique, Université Montpellier II - France

Groupe de Biostatistique et d'Analyse des Systèmes, ENSAM-INRA-UM II

e-mail : jfd@helios.ensam.inra.fr

Abstract:

We present a recently devised extension of the linear PLS model to the non-linear additive that we call PLSS, through the transformation of predictors by polynomial spline functions. The classical Least-Squares Splines estimator can be considered as a particular PLSS estimator when the number of carried out PLSS components is large enough. Suitable tuning spline parameters of PLSS allow the user to experiment with a wide range of PLS regression tools, from usual linear and polynomial models towards more flexible local polynomial additive modeling. Due to B-spline basis functions, PLSS models are not very sensitive to extreme values of the predictors in contrast to most component based regressions. This paper aims at presenting the method like a user's guide and a real example of sensory analysis illustrates the performance of PLSS in the presence of outliers and non-linear relationships.

Key words: Additive Models, B-splines, Least-Squares Splines, PLS Methods.

1 Introduction

In many experimental situations one is usually faced with data sets where the ratio of observations to variables is small and the predictors are highly correlated. Such a regression context may cause harm in linear modeling by Ordinary Least-Squares (OLS). To remedy this problem, Partial Least-Squares regression (PLS), Wold H. (1975), Wold S. et al (1983) see also Tenenhaus (1998), is attractive because it builds principal components at the same time as (partial) regressions are processed thus constructing new uncorrelated predictors with a better predictive potential than those from the Principal Component Regression (PCR).

Let $(x^1, \ldots, x^p)$ be a set of predictors related to a set of responses $(y^1, \ldots, y^q)$, all measured on the same n individuals with sample data matrices $\boldsymbol{X}$ $(n \times p)$ and $\boldsymbol{Y}$ $(n \times q)$ whose columns are respectively denoted $\boldsymbol{x}^i$ and $\boldsymbol{y}^j$. These matrices are column centered with respect to $\boldsymbol{D}$, a $n \times n$ diagonal matrix of statistical weights for the observations (by default $\boldsymbol{D} = \frac{1}{n}\boldsymbol{I}_n$) so that, from a geometrical point of view, the covariance is expressed as the $\boldsymbol{D}$-scalar product $cov(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}'\boldsymbol{D}\boldsymbol{y}$.

The linear PLS fit for the jth response

$$\widehat{y}^j_A = \hat{\beta}^{j,1}_A x^1 + \ldots + \hat{\beta}^{j,p}_A x^p \qquad (1)$$

depends on the model dimension A which is the number of the carried out components. When $A = rank(\boldsymbol{X})$, both OLS and PLS models are identical which enlightens about the scope of application for PLS: few observations and/or highly correlated predictors.

A multi-response additive spline model is a fit of the form

$$\widehat{y}^j_A = s^{j,1}_A(x^1) + \ldots + s^{j,p}_A(x^p) \tag{2}$$

where the coordinate function $s^{j,i}_A(x^i)$ is a spline function which is a piecewise polynomial that measures the additive influence of the predictor x^i on the response y^j. A coordinate function may depend on the number A of components if any dimension reduction procedure is used like in PLSS for example.

Actually, PLS through Splines, in short PLSS, (Durand 1997, 1999, 2000) is the standard PLS regression between the response(s) $\boldsymbol{Y}$ and the coding matrix $\boldsymbol{B}$ whose columns are constructed by transforming the predictors through B-spline basis functions

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

This approach mainly differs from that of Durand and Sabatier (1997) in that PLSS uses a singular value decomposition to compute components, namely that of the PLS kernel, instead of a time consuming iterative procedure, see (Durand 1997) for a comparison between these two methods. The price to be payed by PLSS for non-linearity is the increase of the column dimension for the new design matrix $\boldsymbol{B}$. In contrast to the Least-Squares Splines (LSS) regression (Stone 1985), expanding the dimension which is tantamount to increasing the number of predictors, is well supported by PLSS that inherits the advantages of the standard PLS method. It is shown that the LSS model is identical to that of PLSS when the number A of components is the rank of $\boldsymbol{B}$.

An attractive property of the linear space of the so-called regression splines spanned by B-spline basis functions (De Boor 1978), enables the PLSS model to capture linear, polynomial as well as local polynomial relationships: the linear space of classical polynomials on the interval $[a, b]$ is a particular spline space when no point where two polynomials join end to end - such a point is called a knot - does exist. For example, the simplest call to the S-Plus (MathSoft 1996) function *plss* implemented by the author, makes use of polynomials of degree one on $[a, b]$ that produces a linear PLS model. By this way, PLSS provides the user with modeling tools that range from the linear model to local polynomial models, including classical polynomial modeling.

In component based regression methods like PCR and PLS, extreme values of the predictors have a global influence on the components and then on the linear model. Using B-spline basis functions in PLSS is very attractive in this context. These functions with a local support, they vanish outside an interval of knots, warrant that extreme values of the predictors have a local influence on the model. The data set illustrating the capabilities of PLSS is typical of the difficulties encountered by the standard PLS method in such a context.

Section 2 presents some practical properties of the linear space of regression splines. Theoretical aspects of PLSS are developed in Section 3 while the practice of the *plss* function is presented in Section 4. The sensorial analysis of the orange juice data (Durand 2000) is the real example that illustrates the capabilities of PLSS.

2 Regression splines in a nugget

The use of splines introduces a nonparametric character in the regression and data speak for themselves to build the model. The tuning parameters for regression splines are the number K and the location of points where adjacent polynomials join end to end (the "interior knots", in short, the "knots") as well as the degree d of the polynomials.

To transform a continuous variable x whose sample vector $\boldsymbol{x}$ ranges on $[a, b]$, when d and the location of K interior knots lying within $]a, b[$ have been fixed, a spline s belongs to a linear functional space of dimension $r = d + 1 + K$. Then, a spline function can be written

$$s(x) = \sum_{k=1}^{r} \beta_k B_k(x)\,,$$

where $\{B_1, \ldots, B_r\}$ is the set of the most popular basis functions called B-splines (De Boor 1978).

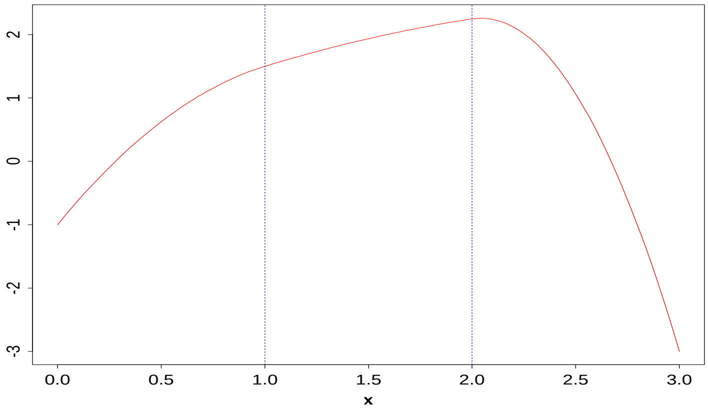

Figure 1. A spline function of degree 2 on [0,3] with spline coefficients (-1, 1, 2, 2.5, -3). Dotted vertical lines indicate the location of knots.

Figure 1 shows the shape of a spline of degree 2 (piecewise quadratic) on the interval [0,3] with knots located at 1 and 2. Coordinate values $\boldsymbol{\beta} = (\beta_1, \dots, \beta_r)$ usually called the spline coefficients, are to be estimated by a regression method.

In contrast to usual polynomial basis functions, $B_k(x) = 0$ outside an interval of knots as shown on Figure 2 (actually, a and b are auxiliary non-interior knots whose roles are not detailed here). As a consequence, one observation x_i (0.5 for example) has a local influence on $s(x_i)$ that depends only on the $d+1$ basis functions (in that case B_1, B_2 and B_3) whose supports encompass this data.

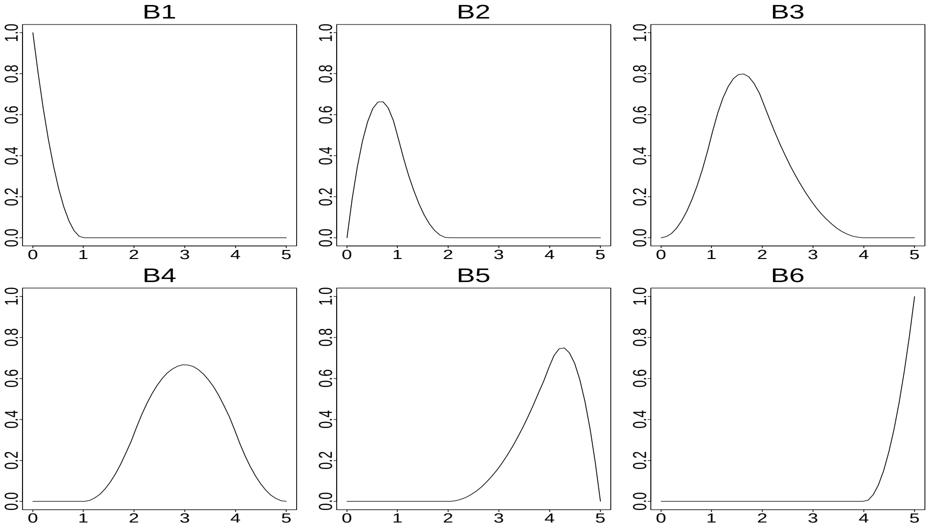

Figure 2. B-splines of degree 2 on $[0, 5]$ with knots at 1,2 and 4.

The counterpart of this advantage is that $s(x) = 0$ outside $[a, b]$ thus indicating that prediction through regression splines is only efficient when the observation to be predicted lies within the data used to build the model.

Furthermore, the B-spline family $\{B_k(.)\}_k$ is a set of fuzzy coding functions on $[a, b]$ with the following properties

$$0 \leq B_k(x) \leq 1\,, \qquad \sum_{k=1}^{r} B_k(x) = 1\,, \qquad \forall x \in [a, b]. \tag{3}$$

Notice that B-splines of degree 0 are piecewise constant binary coding functions: $B_k(x)$ is 1 within an interval of adjacent knots and 0 outside.

When knots are missing ($K = 0$), Figure 3 shows how the corresponding $\{B_k(.)\}_k$ family is a basis for usual polynomials on $[a, b]$.

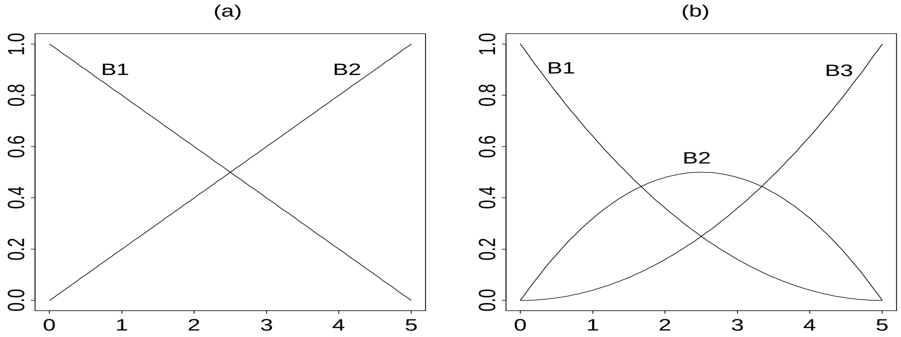

Figure 3. B-spline basis functions without knots for the space of polynomials on $[0, 5]$ with degree 1, (a), and degree 2, (b).

This particular case is important from a practical perspective. Any regression model based on B-spline basis functions, can capture linear as well as polynomial relationships on $[a, b]$ through the use of an empty set of knots.

We refer to De Boor (1978) for the computation of B-spline functions and for more information on the smoothness of $s(x)$ at a knot, that is, on how two adjacent polynomials join end to end, see also (Durand 2000).

In the multivariate context, denote r_i the dimension of the spline space for the predictor x^i and $r = \sum_{i=1}^{p} r_i$ the total dimension. The $n \times r$ matrix

$$\boldsymbol{B} = [\boldsymbol{B}^1 | \dots | \boldsymbol{B}^p] \tag{4}$$

has elements that are the codings of the n observations on the predictors through B-spline functions. More precisely, the kth column in the $n \times r_i$ block $\boldsymbol{B}^i$ is constructed from the coordinates of $\boldsymbol{x}^i$, through their transformation by the corresponding kth B-spline function.

In the matrix version of the additive spline model (2), the jth column of the $n \times q$ model matrix $\widehat{\boldsymbol{Y}}_A$ is the sum of the smoothed sample predictors

$$\widehat{\boldsymbol{y}}^j_A = \sum_{i=1}^{p} \boldsymbol{B}^i \, \widehat{\boldsymbol{\beta}}^{j,i}_A = \sum_{i=1}^{p} s^{j,i}_A(\boldsymbol{x}^i) \, , \tag{5}$$

where the column $\boldsymbol{x}^i$ is transformed by the ith coordinate spline function to get the column vector $s^{j,i}_A(\boldsymbol{x}^i) = \boldsymbol{B}^i \, \widehat{\boldsymbol{\beta}}^{j,i}_A$. The response(s) being centered, coordinate functions in (2) are identifiable if $E[s^{j,i}_A(x^i)] = 0$ which implies that any matrix $\boldsymbol{B}^i$ must be column centered. Then, just as in the well known particular case of binary indicator matrices, the centered matrix $\boldsymbol{B}$ is not of full column rank as a consequence of fuzzy coding properties (3),

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

A natural extension of the linear model to additivity through regression splines is the Least-Squares Splines (LSS) estimator of Stone (1985), see also (Eubank 1988). It consists in replacing the $\boldsymbol{X}$ design by the centered super-coding matrix $\boldsymbol{B}$. One obtains directly an additive model (2), (5), without dimension reduction, by performing the Ordinary Least-Squares regression

$$LSS(\boldsymbol{X}, \boldsymbol{Y}) \equiv OLS(\boldsymbol{B}, \boldsymbol{Y})\,. \tag{6}$$

As a consequence of the remark on the rank of $\boldsymbol{B}$, the LSS estimator is computed by removing the first column in each block $\boldsymbol{B}^i$ to obtain a full column ranked design matrix. However, the LSS estimator does exist only if $\boldsymbol{B}$ has a sufficiently large ratio of observations to variables. An another well known drawback inherited from the OLS model, is that the LSS model is sensitive to multicollinearity in the predictors.

The preceding limits plead on behalf of a method like PLSS whose component based approach firstly solves the problem of the dimension when extending the linear model to the additive, and also provides robust models in the presence of strongly correlated predictors.

3 The PLSS regression

Let us first briefly recall standard notation and algebra of the PLS kernel, having in view the geometrical aspects of the method mainly based on projections. We refer to (Durand, Roman et Vivien, 1998) for the S-Plus user’s guide of the function *pls* implemented by R. Sabatier and the author.

3.1 The PLS model

PLS constructs a sequence of centered and uncorrelated explanatory variables $\{\boldsymbol{t}^1, \ldots, \boldsymbol{t}^A\}$ called the PLS components, to predict the response(s) in a linear fashion. Denote $\boldsymbol{E}_0 = \boldsymbol{X}$ and $\boldsymbol{F}_0 = \boldsymbol{Y}$ the sample data matrices on the predictors and responses respectively whose columns are standardized with respect to statistical weights in the diagonal of $\boldsymbol{D}$. The step k of PLS which constructs the component $\boldsymbol{t}^k$ may be decomposed in two steps presented in Table 1.

Table 1. The PLS kernel

1) Construct $\boldsymbol{t}^k$, for fixed $\boldsymbol{E}_{k-1}$, $\boldsymbol{F}_{k-1}$:
$\boldsymbol{t} = \boldsymbol{E}_{k-1}\boldsymbol{w}, \quad \boldsymbol{u} = \boldsymbol{F}_{k-1}\boldsymbol{c};$
$(\boldsymbol{w}^k, \boldsymbol{c}^k) = arg \max\limits_{\boldsymbol{w}'\boldsymbol{w}=\boldsymbol{c}'\boldsymbol{c}=1} cov(\boldsymbol{t}, \boldsymbol{u});$
$\boldsymbol{t}^k = \boldsymbol{E}_{k-1}\boldsymbol{w}^k, \quad \boldsymbol{u}^k = \boldsymbol{F}_{k-1}\boldsymbol{c}^k.$
2) Update $\boldsymbol{E}_k$ and $\boldsymbol{F}_k$:
$\boldsymbol{E}_k = \boldsymbol{E}_{k-1} - \boldsymbol{P}_{\boldsymbol{t}^k}\boldsymbol{E}_{k-1},$
$\boldsymbol{F}_k = \boldsymbol{F}_{k-1} - \boldsymbol{P}_{\boldsymbol{t}^k}\boldsymbol{F}_{k-1}\,.$

In 1), the objective function to be maximized is the covariance between linear compromises $\boldsymbol{t}$ and $\boldsymbol{u}$ of respectively $\boldsymbol{E}_{k-1}$ and $\boldsymbol{F}_{k-1}$. Note that in the one-response case, $\boldsymbol{u}^1$ is equal to the response $\boldsymbol{Y}$. The solution $(cov(\boldsymbol{t}^k, \boldsymbol{u}^k), \boldsymbol{w}^k, \boldsymbol{c}^k)$ is the triple associated with the largest singular value $cov(\boldsymbol{t}^k, \boldsymbol{u}^k)$ in the singular value decomposition of the covariance matrix $\boldsymbol{E}'_{k-1}\boldsymbol{D}\boldsymbol{F}_{k-1}$ “between” the updated explanatory and response variables.

Step 2) updates the predictors and the responses as the residuals of the regressions on the component $\boldsymbol{t}^k$ computed in 1). The matrix $\boldsymbol{P}_{\boldsymbol{t}^k} = \boldsymbol{t}^k \boldsymbol{t}^{k\prime} \boldsymbol{D}/var(\boldsymbol{t}^k)$ is the $\boldsymbol{D}$-orthogonal projection matrix on the component. Steps 1) and 2) overlap each other when the NIPALS (Nonlinear Iterative PArtial Least-Squares) loop of Wold (1966) is used for an automatic treatment of missing data. The current version 9.6 of the S-Plus *plss* function does not propose this possibility. However, because of the straightforward extension of PLS defined in (11), we conjecture that there is no theoretical difficulty for handling missing data through NIPALS and PLSS.

The preceding discussion shed some light on a particular property of PLS compared to Principal Component Analysis (PCA) that may be considered as a “self-PLS” regression of $\boldsymbol{X}$ onto itself

$$PLS(\boldsymbol{X}, \boldsymbol{Y} = \boldsymbol{X}) \equiv PCA(\boldsymbol{X})\,.$$

In this case, $\boldsymbol{t}^k = \boldsymbol{u}^k$ and the criterion is that of the maximal variance.

Components $\{\boldsymbol{t}^1, \dots, \boldsymbol{t}^A\}$ are centered variables. They are mutually uncorrelated and constitute a $\boldsymbol{D}$-orthogonal basis for a subspace of the linear space $span(\boldsymbol{X})$ spanned by the columns of $\boldsymbol{X}$

$$\boldsymbol{t}^k = \boldsymbol{X}\boldsymbol{a}^k\,, \qquad k = 1, \dots, A. \tag{7}$$

Denote $\boldsymbol{T}_A = [\boldsymbol{t}^1, \dots, \boldsymbol{t}^A]$ the $n \times A$ matrix of the PLS components. The linear PLS model, allows to reconstruct $\boldsymbol{X}$ and to explain $\boldsymbol{Y}$ in terms of the components

$$\boldsymbol{X} = \widehat{\boldsymbol{X}}_A + \boldsymbol{E}_A = \boldsymbol{P}_{\boldsymbol{T}_A}\boldsymbol{X} + \boldsymbol{E}_A$$

$$\boldsymbol{Y} = \widehat{\boldsymbol{Y}}_A + \boldsymbol{F}_A = \boldsymbol{P}_{\boldsymbol{T}_A}\boldsymbol{Y} + \boldsymbol{F}_A\,, \tag{8}$$

where $\boldsymbol{P}_{\boldsymbol{T}_A} = \sum_{k=1}^{A} \boldsymbol{P}_{\boldsymbol{t}^k}$ is the matrix (of rank A) of the $\boldsymbol{D}$-orthogonal projection on $span(\boldsymbol{T}_A)$.

As a consequence, when the model dimension A is equal to the rank of $\boldsymbol{X}$, $\boldsymbol{E}_A = 0$, the linear spaces $span(\boldsymbol{T}_A)$ and $span(\boldsymbol{X})$ coincide and the PLS model (8) is the usual linear OLS model

$$PLS(\boldsymbol{X}, \boldsymbol{Y}) \equiv OLS(\boldsymbol{X}, \boldsymbol{Y}) \qquad \text{when} \qquad A = rank(\boldsymbol{X})\,. \tag{9}$$

Equations (7) and (8) lead to the matrix form of the PLS model (1)

$$\widehat{\boldsymbol{y}}^j_A = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^j_A, \qquad j = 1, \ldots, q, \tag{10}$$

where the vector $\widehat{\boldsymbol{\beta}}^j_A$ of PLS coefficients depends on the number A of components. This tuning parameter is usually estimated by a cross-validation procedure.

3.2 The PLSS model

The additive spline PLSS model is obtained by the PLS regression of the standardized response(s) $\boldsymbol{Y}$ on centered spline transformations $\boldsymbol{B}$ of the standardized explanatory variables $\boldsymbol{X}$

$$PLSS(\boldsymbol{X}, \boldsymbol{Y}) \equiv PLS(\boldsymbol{B}, \boldsymbol{Y})\,. \tag{11}$$

As a consequence of (6), (9) and (11), both PLSS and LSS models are identical when the model dimension A is the rank of $\boldsymbol{B}$.

$$PLSS(\boldsymbol{X}, \boldsymbol{Y}) \equiv OLS(\boldsymbol{B}, \boldsymbol{Y}) \equiv LSS(\boldsymbol{X}, \boldsymbol{Y}) \quad \text{when} \quad A = rank(\boldsymbol{B})\,.$$

Then, PLSS ranks among the family of additive spline models (Hastie & Tibshirani 1990) as a robust method against multicollinearity.

The fact that $\boldsymbol{B}$ is not of full column rank, is an algebraic artifact with no statistical consequence since regressions are made on uncorrelated components $\{\boldsymbol{t}^k\}$ built from $\boldsymbol{B}$ and $\boldsymbol{Y}$. Replacing $\boldsymbol{X}$ by $\boldsymbol{B}$ in (7), a component $\boldsymbol{t}^k$ which is now a linear compromise of $\boldsymbol{B}$, is additively modeled in the predictors. Regrouping the terms, we get

$$\boldsymbol{t}^k = \boldsymbol{B}\boldsymbol{a}^k = \sum_{i=1}^{p} \boldsymbol{B}^i \boldsymbol{a}_i^k = \sum_{i=1}^{p} \varphi_i^k(\boldsymbol{x}^i)\,, \qquad k = 1, \ldots, A. \tag{12}$$

Coordinate function plots $(x^i, \varphi_i^k(x^i))$ are used for interpreting the additive influence of data from predictors on a component and practitioners of the classical PLS regression will have to get accustomed to analyzing non-linearly scatterplots of observations as in Section 4.4.

PLSS model (5) is obtained in the same way for the responses when $\boldsymbol{X}$ is replaced by $\boldsymbol{B}$ in (10). Regrouping the terms leads to

$$\widehat{\boldsymbol{y}}_A^j = \boldsymbol{B}\widehat{\boldsymbol{\beta}}_A^j = \sum_{i=1}^{p} \boldsymbol{B}^i \, \widehat{\boldsymbol{\beta}}_A^{j,i} = \sum_{i=1}^{p} s_A^{j,i}(\boldsymbol{x}^i)\,. \tag{5}$$

The non-linear additive influence of the predictors on the jth response, is interpreted as for components by looking at the most significant coordinate function plots $(x^i, s_A^{j,i}(x^i))$. The next section indicates how to select and order the predictors according to the range of $s_A^{j,i}(\boldsymbol{x}^i)$. Other selection of variables can be done, for example, by grouping the predictors that present the same coordinate function shape.

To end this introduction to PLSS, let us recall the tuning parameters of the model:

- the type a the spline space for each predictor (degree, number and location of knots),
- the number of components.

4 The PLSS practice

To capture non-linear relationships, no automatic procedure for selecting optimal spline parameters is at disposal in the current version 9.6 of the S-Plus function denoted *plss* and more extensively detailed in (Durand 2000). The user has to decide on the type of splines and two strategies can be followed to that aim.

The ascending strategy consists in increasing progressively the degree and the number of knots. By default, *plss* proposes, for all the predictors, splines of degree 1 without knots, that is, linear polynomials that lead to a linear PLS model. This option constitutes a reasonable departure point in the quest of an ideal additive model. Then, increasing the degree while keeping at zero the number of knots permits to explore the usual polynomial models. Finally, more flexible local polynomial models may be found sharing the domain of each predictor in joint intervals and choosing the splitting points as knots for spline transformations.

The main heuristic rule for knots can be formulated as: "adding a knot increases the local flexibility of the spline and then, the freedom of fitting the data in this area".

In contrast, the descending strategy starts with a high degree, three for example, and more knots than necessary (a sufficiently large number depending on both the degree and the data); then, superfluous knots are removed and the degree is decreased as much as possible.

Looking at the evolution of coordinate function shapes for each predictor (when p is not too large) while examining both sequences of values for the goodness-of-fit $R^2(A)$ as well as for the Prediction Residual Sum of Squares $PRESS(A)$, can help to decide when the selected strategy is to be stopped. The goodness-of-fit criterion $R^2(A)$ used in *plss* is the proportion of the total $\boldsymbol{Y}$-variance accounted for by A components which is an increasing function of A

$$R^2(A) = \frac{1}{q}\sum_{j=1}^{q} R^2(\boldsymbol{y}^j, span(\boldsymbol{t}^1, \dots, \boldsymbol{t}^A))\,. \tag{13}$$

When candidates for splines have been selected, the construction of $PRESS(A)$ by the PLSS cross-validation procedure is identical to that of the usual PLS regression. For example, in the leave-one-out case, the ith observation is removed, that is, the ith row of $\boldsymbol{B}$ and $\boldsymbol{Y}$, and the $r \times q$ matrix $\hat{\boldsymbol{\beta}}_A^{(-i)}$ of the PLSS model with A components, is based on the remaining observations. Performing this procedure for $i = 1, \dots n$, the mean squared error of prediction $PRESS^j(A)$ is computed for each response j and the total $PRESS$ is a function of A (expected to be firstly decreasing)

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

There is no explicit formula for stopping both ascending and descending strategies and the building-model stage consists in locating knots that achieve a balance between "thriftiness" (of dimensions r and A) and "goodness" (of fit and prediction) for the model candidates. To avoid overfitting, one has to carefully look for parsimonious PLSS models with better values for both R^2 and $PRESS$ criteria as a result of possibly smaller dimensions of both splines and components.

The ascending strategy followed on the data presented in the next section, presents the advantage of its starting point that allows to know whether the linear PLS model is valid or not. In this example, the answer will be negative (improvements can be tried through non-linear analyses) mainly due to outliers and apparent non-linear structures in the scatterplots of Figure 4.

4.1 The orange juice data

Our real example is a data set from sensometrics, that consists of twenty-four observations (orange juices) and eleven variables analyzed in (Durand 2000). More precisely, ten predictors characterizing the mineralogical properties of the juices have been measured to explain one sensory response. The twenty-four orange juices which all have identical quantity of fruit and same glucide content, have been named by the capital letters "A", ..., "X" for confidentiality. The ten predictors are the conductivity $COND$, the eight mineralogical characters SiO_2, Na, K, Ca, Mg, Cl, SO_4, HCO_3 and their Sum. Predictors $COND$,

Ca, Mg, SO_4 and Sum are strongly correlated as well as SiO_2 and K. The response is the sensory descriptor $Heavy$ whose protocol of measurements is not revealed for confidentiality.

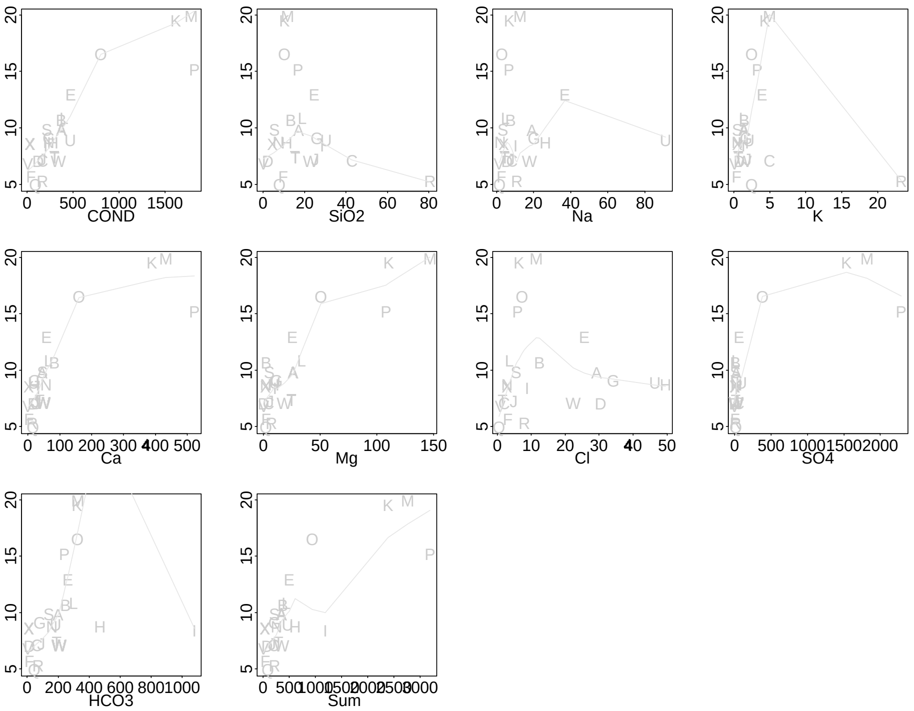

Figure 4. Bivariate plots between $Heavy$ and the predictors. The observations are smoothed by the S-Plus function “lowess”.

Moreover, Figure 4 displaying bivariate plots of the response $Heavy$ with all the predictors, shows some non-linear structures as well as extreme values in the predictors that sometimes may be considered as outliers. The trend of the bivariate relationships is indicated by the S-Plus linear smoother “lowess”, see (Venables and Ripley 1994).

4.2 The default option for *plss*: linear PLS

The simplest call to the function *plss*, namely $plss(X, Y)$, makes use of splines without knots with degree 1 for all the predictors thus proposing the PLS regression on linear basis functions as displayed in Figure 3 (a). However the linear PLSS model differs from standard PLS in that the coefficients of the model are not directly interpretable since they are associated to the linear basis functions and not to the explanatory variables.

Selecting the dimension of the model is a crucial choice in PLS. For this, the first indication is the goodness-of-fit criterion $R^2(A)$. The second is given by the evolution of the $PRESS$ values with different model dimensions.

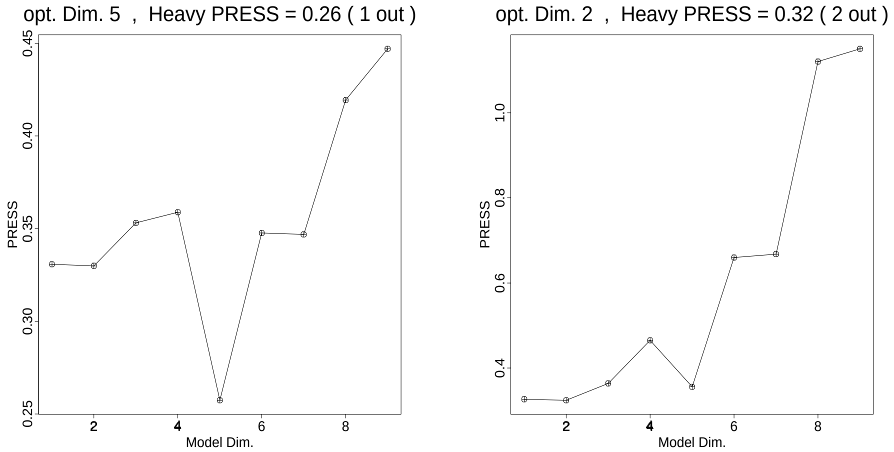

Figure 5. Different $PRESS$ of the PLSS linear model according to model dimensions when 1 or 2 samples are left at a time.

The R^2 between $Heavy$ and $\boldsymbol{t}^1$ is equal to 0.754 and Figure 5 displays the evolution of the $PRESS$ with successive dimensions A, when one or two observations are left and predicted. To control the influence of outliers, some "two out" cross-validations were made on permuted samples. As a result of many tries, our decision was to select one component for the model.

Figure 6 presents the model for *Heavy* with one component in an additive spline vision of the linear model. On a coordinate function plot, the horizontal axis represents the standardized predictor, while its centered spline transformation is displayed on the vertical axis.

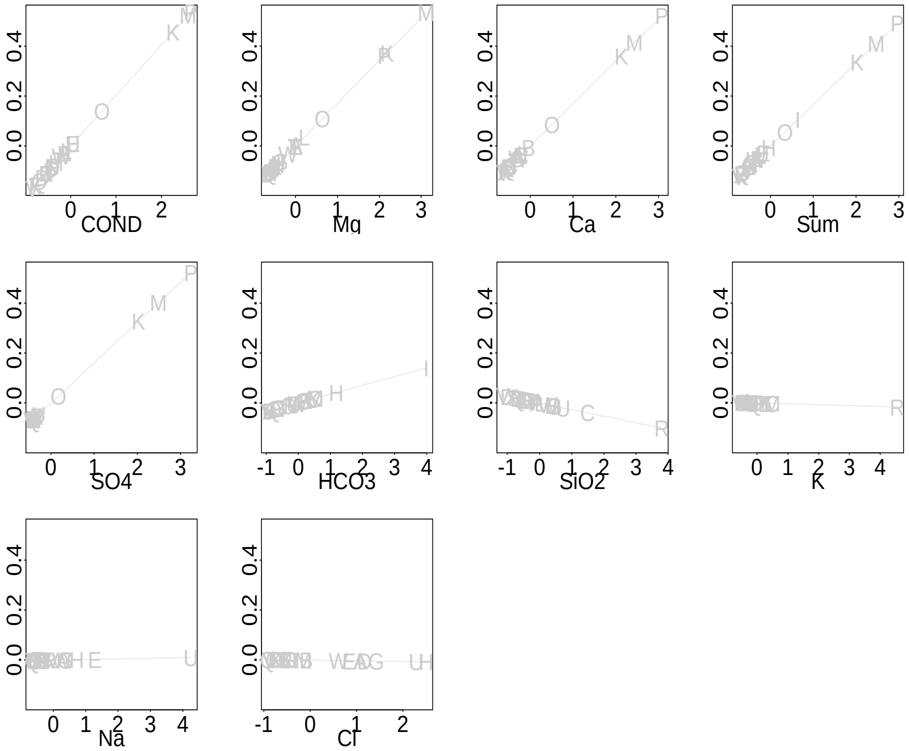

Figure 6. Linear PLSS model for *Heavy* with 1 component.

Plots are ordered from left to right and from up to down according to the range of the transformed data. Predictors are then ordered through their decreasing influence on the response. For the five first variables, the positive slopes corroborate the trends of the corresponding bivariate plots of Figure 4. The last five predictors with no influence, are those for which outliers are present in the bivariate scatterplots.

4.3 Increasing the degree and the knots

We selected the degree 2 and omit to present the exploration of usual quadratic models through the call $plss(X, Y, degree = 2)$. Note that each predictor can be transformed with an individual degree when *degree* is a vector of p positive integers.

The first and swift way of managing the knots is to locate them either at quantiles or at equally spaced positions within the domain of a predictor. For this, their number is controlled by the integer parameter *knots* (by default 0) associated with the boolean parameter *equiknots* (by default F for knots at quantiles). For instance, $plss(X, Y, degree = 2, knots = 3)$ supplies coordinate spline functions of degree 2 with knots located at quartiles for all the predictors. These two parameters can also be vectors of p individual choices. Without a priori information on the data, this strategy is worth considering especially in the presence of a large number of predictors, as a first exploration before more suitable future investigations.

The second strategy based on some knowledge of the data, is to decide that the predictors need an individual specific control of their knots through the use of the parameter *listknots* which is a list (missing by default) of p vectors candidate for knots' locations

$$plss(X, Y, degree = 2, listknots = list(knots_1, \ldots, knots_p)).$$

Table 2. Selected knots for the predictors

$COND$	SiO_2	Na	K	Ca	Mg	Cl	SO_4	HCO_3	Sum
400	10	10	2.5	160	40	4	400	100	600
1600	20	40	5	400	110	11	1700	300	2600
	40					30		500	

Both strategies for knots have been tried on the orange juice data but the second appeared more useful to isolate extreme values of the predictors and restrict their influence in a more local way. Bivariate scatterplots of Figure 4 provide some informations about the number and the location of possible knots and Table 2 presents the candidates retained for knots.

Note however that the difficulty of choosing knots in the domain of the predictors is increased when two responses, or more, are simultaneously predicted. For this reason we recommend to perform multi-response PLSS modeling when the responses are strongly correlated. In that case, bivariate plots between the responses and one explanatory variable supply similar hints for locating knots and selecting the degree of the spline transformation.

The convex shape of $PRESS(A)$ shown in Figure 7, and the evolution of the R^2 (0.854 and 0.917) between $Heavy$ and the subspaces $span(\boldsymbol{t}^1)$, $span(\boldsymbol{t}^1, \boldsymbol{t}^2)$ respectively, indicate that the optimal number of components is $A = 2$. The goodness-of-prediction is greatly improved compared with that of the linear model (0.15 against 0.33 for the PRESS values).

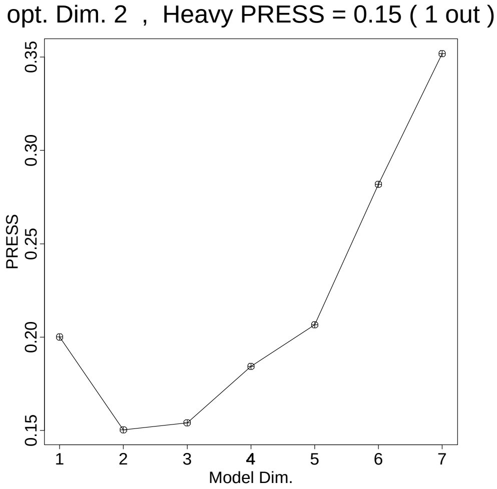

Figure 7. $PRESS$ of the PLSS model with the option "1 out".

Coordinate function plots of the model shown Figure 8, can be examined in the same way of those in Figure 6. Predictors have been ordered according to their decreasing additive influence on $Heavy$. For the five predictors of main influence, SO_4, Ca, Mg, $COND$ and Sum, the shapes of the splines corroborate the underlying non-linear bivariate structures displayed Figure 4. In the second group of the last five predictors, outliers I and R do not perturb the role of the other orange juices in the construction of coordinate functions for HCO_3 and K. This is mainly due to the location of knots that isolate these data. Actually, the gap of influence between the two groups of predictors excessively marked in the linear model of Figure 6, is significantly reduced in this local polynomial PLSS model.

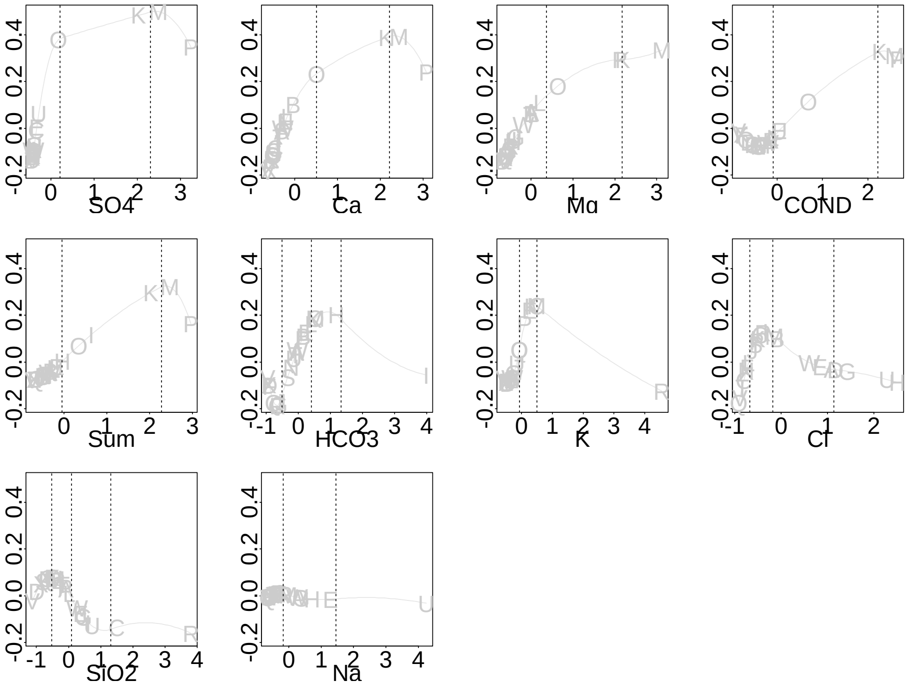

Figure 8. Additive spline model for *Heavy* with 2 components. The degree is 2 and the dotted vertical lines indicate the location of knots.

4.4 A non-linear look at data

PLSS principal component scatterplots $(\boldsymbol{t}^i, \boldsymbol{t}^j)$ are used to look at data having in view that components are built to explain the responses in the same linear way as PLS does. Figure 9 (a) displays the 24 products in the $(\boldsymbol{t}^1, \boldsymbol{t}^2)$ plot and Figure 9 (b), plotting correlations between *Heavy* and the first two components, shows that the response is strongly correlated with $\boldsymbol{t}^1$.

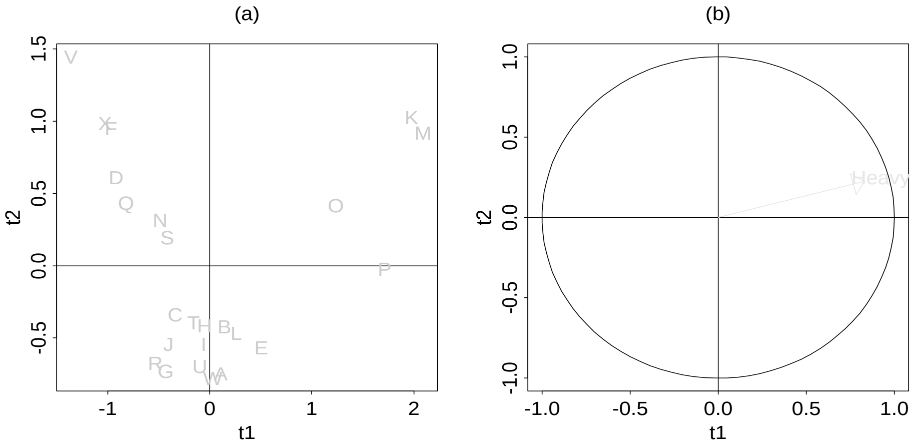

Figure 9. First PLSS principal component scatterplot, (a), and correlations between *Heavy* and the two first components, (b).

Clearly, orange juices M, K, P and O present the greatest *Heavy* character as well as E, L, B and A but to a lesser degree. This second group of cases cannot be detected in the corresponding scatterplot of the linear model not presented here. The difference between linear and non-linear PLSS data analysis is to be found in the interpretation of the components by the predictors. More precisely, one has to look at data through non-linear coordinate spline functions expressed in (12).

In Figure 10 displaying the coordinate function plots for $\boldsymbol{t}^1$, predictors are classified in decreasing order as in Figures 6 and 8, according to the range of the transformed data. The four preponderant orange juices M, K, P and O present large transformed values (above the mean) for the main variables SO_4, Ca, Mg, Sum, K and $COND$. Note however that P is less influent than M and K through this non-linear model while the contrary holds in the linear model. The variables K, HCO_3

and Cl with no influence in the linear model due to the presence of extreme data, are now appearing as more influent. This partially explains the fact that juices E, L, B and A are revealed in Figure 9 (a) as presenting values above the mean for the character $Heavy$.

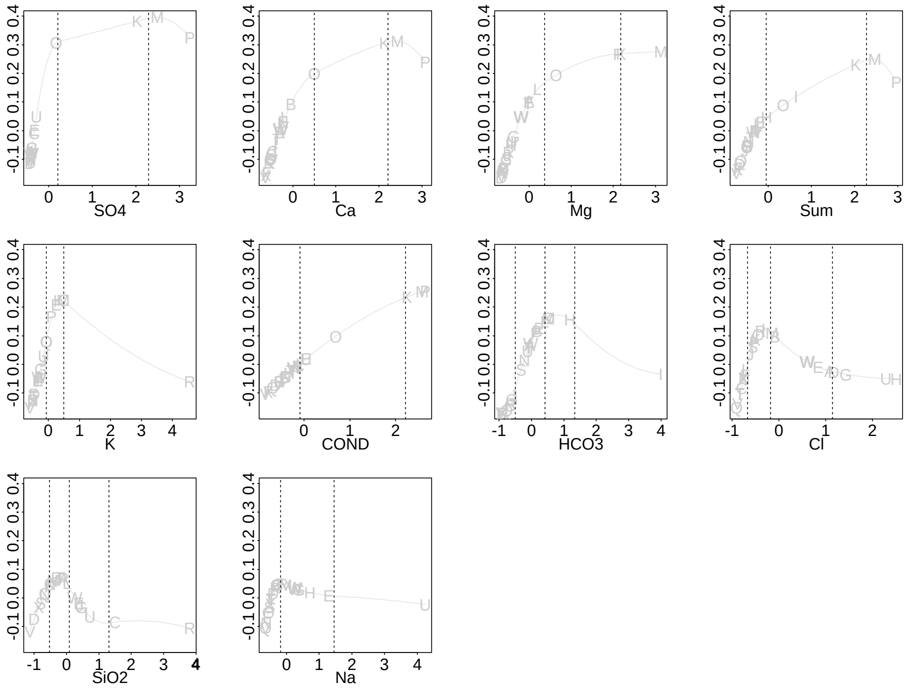

Figure 10. Data and coordinate functions for $\boldsymbol{t}^1$. The degree is 2 and the dotted vertical lines indicate the location of knots.

5 Conclusions

Multi-response PLSS models were not experimented here and we refer to (Durand 2000) for a detailed analysis of the complete orange juice data set. We have neither presented in this paper the use of B-splines in some specific cases. In the discriminant analysis context, pre-processing B-spline transformations of a variable $\boldsymbol{y}$ to obtain a fuzzy (binary, with B-splines of degree 0) indicator matrix $\boldsymbol{B}_y$ of groups, produces through $PLSS(\boldsymbol{X}, \boldsymbol{Y} = \boldsymbol{B}_y)$ a non-linear additive discriminant model. On the other hand, using B-splines of degree 0 in PLSS leads to an ANOVA PLS model, see (H. Martens & M. Martens 1999) for assessing the significance of model coefficients.

Only main features of PLSS were presented and the performance of the method was illustrated on a data set from sensometrics. PLSS is a straightforward extension of standard PLS that constructs additive spline models. This method inherits from PLS the traditional robustness of the models against both scarcity of data and multicollinearity of the predictors.

In contrast to most component based methods prone to outliers, the use of B-spline basis functions protect in some way the models from an excessive influence of extreme data.

The technique involved does not automatically estimates the optimal transformations of the predictors. However, experimenting with the tuning spline parameters allows the PLSS user to explore a wide class of polynomial PLS models from the usual linear to the more flexible local polynomials. Practitioners of the standard PLS method will easily become acquainted with this new approach not so far from the traditional way of predicting and analyzing multivariate data.

Bibliography

De Boor C. (1978). *A practical guide to splines*, Berlin: Springer-Verlag.

Durand J.F. (1997). *Additive Modeling of multivariate data by spline functions*, Rapport de Recherche 97-04, Groupe de Biostatistique et d'Analyse des Systèmes, ENSAM-INRA-UM II.

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. (2000). *La régression Partial Least Squares Spline - PLSS - guide d'utilisation sous S-Plus*, Rapport de Recherche 00-06, Groupe de Biostatistique et d'Analyse des Systèmes, ENSAM-INRA-UM II.

Durand J.F., Roman S. & Vivien M. (1998). *Guide d'utilisation de la régression Partial Least Squares Linéaire sous S-Plus*, Rapport de Recherche 98-06, Groupe de Biostatistique et d'Analyse des Systèmes, ENSAM-INRA-UM II.

Durand J.F., & Sabatier R. (1997). Additive Splines for PLS regression, *Journal of the American Statistical Association*, Vol. 92, 440, 1546-1554.

Eubank R. L. (1988). *Spline smoothing and nonparametric regression*, New York: Dekker.

Hastie T., & Tibshirani R. (1990). *Generalized additive models*, London: Chapman & Hall.

MARTENS H., & MARTENS M. (1999). Validation of PLS Regression models in sensory science by extended cross-validation, *Les Méthodes PLS, Symposium International PLS'99*, Tenenhaus, M. et Morineau, A. (Eds), CISIA-CERESTA, 149-182.

MATHSOFT (1996). *S-plus Version 3-4 for Unix Supplement.* Data Analysis Products Division, Mathsoft, Seattle.

STONE, C. J. (1985). Additive regression and other nonparametric models, *Annals of Statistics*, 13, 689-705.

TENENHAUS, M. (1998). *La régression PLS, théorie et pratique*, Paris: Technip.

VENABLES W.N., & RIPLEY B.D. (1994). *Modern Applied Statistics with S-Plus*, New York: Springer-Verlag.

WOLD, H. (1966). Estimation of principal components and related methods by iterative least squares, in *Multivariate Analysis*, ed. P.R. Krishnaiah. New-York: 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., & WOLD H. (1983). The multivariate calibration problem in chemistry solved by PLS method, *Proc. Conf. Matrix Pencils.* Ruhe, A. and Kagstrom, B. (Eds), Lecture notes in mathematics, Heidelberg: Springer Verlag, 286-293.