Research Article

JOURNAL OF CHEMOMETRICS

Received: 18 February 2009, Revised: 4 August 2009, Accepted: 14 September 2009, Published online in Wiley Interscience: 17 November 2009

(www.interscience.wiley.com) DOI: 10.1002/cem.1260

Model building in multivariate additive partial least squares splines via the GCV criterion

R. Lombardo[a]*, J. F. Durand[b] and R. De Veaux[c]

In the literature, much effort has been put into modeling dependence among variables and their interactions through nonlinear transformations of predictive variables. In this paper, we propose a nonlinear generalization of Partial Least Squares (PLS) using multivariate additive splines. We discuss the advantages and drawbacks of the proposed model, building it via the generalized cross validation criterion (GCV) criterion, and show its performance on a real dataset and on simulated datasets in comparison to other methods based on splines. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: multivariate B-splines; variable interactions; partial least squares via splines; analysis of variance decomposition; GCV criterion.

1. INTRODUCTION

The fitting of noisy data, especially in high dimensions, with many correlated predictors is a central topic in the literature of predictive modeling. Partial Least Squares (PLS), whose origin dates back to the 1960s [1,2], became a popular method for the linear prediction and modeling of high dimensional data, especially in the field of chemometrics. Many papers have presented nonlinear generalizations of PLS either by replacing the standard regression functions with nonlinear functions (e.g. [3]), or by using spline functions to transform the predictors [4–6].

In this paper, we discuss the potential of generalizing PLS with multivariate additive spline transformations, a method we call Multivariate Additive Partial Least Squares Splines (MAPLSS). A simple version of this method was already proposed by the first two authors [6]. In the present study, to overcome the problem of high computational costs due to the use of multivariate spline transformations, we propose a new MAPLSS model building stage via the generalized cross validation criterion (GCV) [7], introducing a backward deletion procedure to prune the model back. Furthermore via simulation studies, we compare the performance of MAPLSS to known nonlinear models, such as BRUTO [8] and Multivariate Adaptive Regression Splines (MARS) [9], in terms of their accuracy; via real studies, we present MAPLSS as a component regression based method and some new explanatory tools of the component plots will be given.

This multi-response regression method has the ability to: (1) summarize a set of predictive noisy measurements on several collinear variables, (2) use different types of predictor variables, (3) automatically deal with outliers and missing data and last but not the least, (4) include multivariate variable interactions. The MAPLSS model can be represented in a form that separately identifies the contributions of each variable and those associated with the different multivariate interactions via an ANOVA decomposition. For the sake of low computational cost, the GCV has been introduced in the model building stage.

The MAPLSS model involves both univariate [5] and multivariate B-spline transformations. It is additive because each fitted response is a sum of transformation functions of both main and interaction effects of each predictor variable. Using the B-splines (basis splines) as the transforming functions, both the chosen predictors and their interactions will be incorporated into the model. Interactions are simply computed by the tensor product of main effect B-splines. Thus, a variable interaction may be introduced into the model by defining a new predictor in the design matrix. The computational price to be paid by MAPLSS, through tensor products of B-spline functions, is exactly proportional to expanding the column dimension of the new design matrix. In MAPLSS, we focus on the problem of selecting the most relevant main and interaction variables via the GCV criterion. The computational procedure will imply a series of forward/backward phases producing a sequence of models.

The paper is structured as follows: piecewise polynomials, and in particular univariate and multivariate regression splines, together with some extensions toward multivariate additive models, such as BRUTO and MARS are briefly reviewed in Section 2. Section 3 reviews some of the properties of both PLS and PLS via Splines (PLSS; [5]). In Section 4, we illustrate MAPLSS through the ANOVA decomposition model. Details of the computational procedure of MAPLSS are illustrated presenting the GCV criterion to validate the model. In Section 5, some

* *Correspondence to: R. Lombardo, Economics Faculty, Second University of Naples, Capua, CE 81043, Italy.*
E-mail: rosaria.lombardo@unina2.it

a *R. Lombardo*
Economics Faculty, Second University of Naples, Capua, CE 81043, Italy

b *J. F. Durand*
Montpellier II University, cedex 5 Montpellier 34095, France

c *R. De Veaux*
Williams College, Williamstown, MA, 01267, USA

applications to simulated datasets illustrate the capabilities of MAPLSS with respect to two known techniques, BRUTO and MARS, furthermore an example on a real dataset fully describes the richness of MAPLSS output.

2. UNIVARIATE AND MULTIVARIATE REGRESSION SPLINES

Let $x = (x^1, \ldots, x^p)$ be a set of predictors related to a set of responses $y = (y^1, \ldots, y^q)$ all supposed to be standardized and measured on the same n individuals with the same statistical weight $1/n$. Let us denote by $\mathbf{X}$ and $\mathbf{Y}$ the sample data matrices whose current columns are $\mathbf{x}^i$ and $\mathbf{y}^j$, respectively. To introduce ideas, we first review some results on multivariate regression splines. Regression splines are piecewise polynomials whose coefficients are computed according to a regression model [18,19]. In the following we discuss the univariate and bivariate cases and some extensions toward multivariate additive models like BRUTO and MARS.

2.1. The univariate and bivariate cases

Spline functions of a variable $x \in IR$, denoted $s(.)$ are piecewise polynomials of degree d on the open interval (x_-, x_+) that agree at K points $\xi_1, \ldots, \xi_K$ called knots. A set of $r = d + 1 + K$ basis functions called B-splines $B_l(.)$ is used to represent any spline as a linear combination

$$s(x, \beta) = \sum_{l=1}^{r} \beta_l B_l(x)$$

Here $\beta = (\beta_1, \ldots, \beta_r)$ is the vector of spline coefficients computed via regression of $y \in IR$ on the $B_l(.)$

$$\hat{y} = s(x, \hat{\beta}) = \sum_{l=1}^{r} \hat{\beta}_l B_l(x) \tag{1}$$

To estimate β we need to construct the $n \times r$ centered design matrix $\mathbf{B}$ which is the coding matrix of the sample $\mathbf{x}$ through the B-spline family. Besides the fact that they are numerically well conditioned and easy to compute by recursion formulas (see Reference [10,11] for computational and mathematical details) B-splines are attractive because they vanish outside an interval called their support which contains their knots. This vanishing property provides protection against extrapolation for extreme values of x. Moreover, B-splines may be considered as a set of fuzzy coding functions since $0 \leq B_l(x) \leq 1$, and

$$\sum_{l=1}^{r} B_l(x) = 1, \text{ for } x \in [x_-, x_+] \tag{2}$$

Due to relation (2), the column centered matrix $\mathbf{B}$ is not of full column rank. More precisely we have

$$\text{rank } \mathbf{B} \leq \min(n - 1, r - 1)$$

A real problem using B-splines is related to the selection of the knots (location and number). In practice, a few well-located knots often suffices, but optimizing their placement and number in an automatic way is a difficult problem [12–14] whose solution is beyond the scope of this paper. To keep the problem simpler, we will assume that the degree of the B-spline and the sequence of knots is fixed after an exploration of the data. An ascending or descending strategy, increasing or decreasing progressively the degree and the number of knots, can be followed in practice [5]. To include bivariate interactions in our model, we introduce the tensor product of two B-spline families. Consider $x \in IR^2$, and two sets of basis functions: $B_j^1(x^1)$ $j = 1, \ldots, r_1$, for representing functions of coordinate x^1, and $B_k^2(x^2)$, $k = 1, \ldots, r_2$ for coordinate x^2. Then the $r_1 \times r_2$ dimensional tensor product basis defined by

$$B_{j,k}(x^1, x^2) = B_j^1(x^1) B_k^2(x^2), \; j = 1, \ldots, r_1, \; k = 1, \ldots, r_2,$$

allows us to represent a two-dimensional spline function by

$$s_{1,2}(x^1, x^2, \beta) = \sum_{j=1}^{r_1} \sum_{k=1}^{r_2} \beta_{j,k} B_{j,k}(x^1, x^2)$$

A natural extension of model (1) is the bivariate ANOVA spline decomposition

$$\hat{y} = s_1(x^1, \hat{\beta}_1) + s_2(x^2, \hat{\beta}_2) + s_{1,2}(x^1, x^2, \hat{\beta}_{1,2})$$

To estimate the spline coefficients, the $n \times (r_1 + r_2 + r_1 r_2)$ centered super-matrix $\mathbf{B} = [\mathbf{B}^1 | \mathbf{B}^2 | \mathbf{B}^{1,2}]$ is composed of three blocks. The first two are the coding matrices obtained by univariate spline transformations of x^1 and x^2, respectively. The columns of the third block are coordinate-by-coordinate products of the columns of the first two.

In the more general setting of $p > 2$ dimensions, adding new interactions of different orders (among three, four, etc. variables) will permit us to generalize the model to the multivariate case as illustrated in the following. Note however that the column-dimension of the design matrix $\mathbf{B}$ grows exponentially fast, a fact that we will consider during model selection.

2.2. Some extensions toward multivariate additive regression spline models

The real strengths of the adaptive spline methodology, like TURBO [15], and BRUTO (inspired by TURBO, [8]) lie in its ability to both select which terms to include and the amount of smoothing for those included in an efficient way. The BRUTO algorithm combines backfitting and smoothing parameter selection, but considers only main variables and not their interactions. In his MARS (Multivariate adaptive regression spline) model, Friedman [9] generalized the Least Squares Spline (LSS) model [18,19] and BRUTO algorithm by including interaction terms. In MARS, the tuning parameters (number of predictors, spline degree, placement of knots) are chosen adaptively as in BRUTO. The particular spline functions used are truncated linear functions, which are optionally replaced by cubic functions at the last stage, after all the functions have been chosen and fit. The result is that MARS is a flexible non-parametric regression model that attempts to meet the objectives of optimal spline parameters and optimal detection of predictor and their interactions. The model can be written in the form

$$\hat{y} = \sum_{i \in K_1} f_i(x^i) + \sum_{(i,j) \in K_2} f_{i,j}(x^i, x^j) + \ldots \tag{3}$$

where the variable sets denoted by K_1, K_2, etc, involving respectively main effects, bivariate interactions, etc., are automatically selected by the MARS algorithm. Like the LSS model, however, when two knots are close together, the truncated linear basis functions can result in ill-conditioned design matrices. As in LSS, the coefficients are estimated by minimizing the residual sum of squares, but the real art of MARS (like for the BRUTO methodology) is in the model building stage. Starting with only a constant function and all basis functions as candidates, the forward phase constructs a large model that first overfits the data and then, subsequently, a backward deletion procedure is applied to prune the model back. Details can be found in Friedman [9]. To estimate the number of basis functions to include in the model, BRUTO and MARS use the Generalized Cross Validation criterion, GCV, see Reference [15], which presents attractive computational properties.

3. MULTIVARIATE ADDITIVE PARTIAL LEAST SQUARES SPLINES

Before introducing the extension to PLS regression [1], we first review some of the properties of PLS and PLSS [5].

3.1. Quick review of PLS

Although proposed as early as 1966, PLS has been primarily promoted in the chemometrics literature as an alternative to ordinary least squares regression, especially when the design matrix is of much smaller rank than the number of predictors. Like principal component regression, PLS regression can be viewed as a projection of response variables **Y** on linear combinations of the original predictors **X**. The resulting transformed predictors are called latent structures or latent variables. In particular, PLS chooses the latent variables as a series of orthogonal linear combinations (under a suitable constraint) that have maximal covariance with linear combinations of **Y**. PLS constructs a sequence of centered and uncorrelated exploratory variables, i.e. the PLS (latent) components $(\mathbf{t}^1, \ldots, \mathbf{t}^A)$. The number A of the retained latent variables, also called the model dimension, is usually estimated by cross-validation (CV). Because the components are linear combinations of the original predictors **X**, we can write the linear PLS model for the response j as the following expression.

$$\hat{y}^j(A) = \sum_{i=1}^{p} \hat{\beta}_i^j(A) x^i \tag{4}$$

Two particular properties make the PLS attractive and establish a link between the geometrical data analysis and the usual regression. First, when $A = \text{rank } \mathbf{X}$,

$$PLS(\mathbf{X}, \mathbf{Y}) \equiv \{OLS(\mathbf{X}, \mathbf{Y}^j)\}_{j=1,\ldots,q}$$

if the OLS regression exists.

Second, the principal component analysis, PCA, of **X** can be viewed as the 'self-PLS' regression of **X** onto itself,

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

3.2. PLSS, an extension toward pure additive models

PLSS is simply the application of linear PLS regression of **Y** onto the centered coding matrix $\mathbf{B} = [\mathbf{B}^1 | \ldots | \mathbf{B}^p]$. It is thus a purely additive model that depends on the dimension A which, in turn, depends on tuning parameters as well as the spline parameters (degree, number and location of knots). The PLSS model, for the jth response can be written as

$$\hat{y}^j(A) = s_1(x^1, \hat{\beta}_1^j(A)) + \ldots + s_p(x^p, \hat{\beta}_p^j(A)) \tag{5}$$

When the dimension A is equal to the rank of **B**, then PLSS is identical to the usual Least-Squares Splines estimator, if it exists,

$$\text{PLSS}(\mathbf{X}, \mathbf{Y}) \equiv \{\text{LSS}(\mathbf{X}, \mathbf{Y}^j)\}_{j=1,\ldots,q} \quad \text{when } A = \text{rank } (\mathbf{B})$$

Furthermore, comparing PLSS with the Non-Linear Principal Component Analysis (NLPCA), see Reference [16], we can say that NLPCA can be considered as the 'self-PLSS' of **X** onto itself

$$\text{PLSS}(\mathbf{X}, \mathbf{Y} = \mathbf{X}) \equiv NLPCA(\mathbf{X})$$

Because the predictors are standardized, a simple way of ordering the predictors with respect to their decreasing influence on $\hat{y}^j(A)$ is to use as a criterion, the range of the $s_i(\mathbf{x}^i, \hat{\beta}_i^j(\mathbf{A}))$ values of the transformed sample $\mathbf{x}^i$. One can also use that criterion to prune the model, by eliminating the predictors of low influence. In the multi-response case, it is important to point out that only effects that are small in relation to all responses are removed. To stop the pruning process, the GCV criterion presented in Section 4 is used to obtain a more parsimonious model (5) resulting in better out-of-sample predictions.

In order to preserve the advantages of PLSS while including interaction terms in the model, the next section introduces MAPLSS that proposes models based on the ANOVA spline decomposition in the same way as MARS does.

3.3. The MAPLSS model

The price to be paid for incorporating interactions of high degree grows rapidly with the number of predictors. Like MARS, MAPLSS starts the model building phase by proposing models with a large number of predictors. Because MARS is based on the ordinary least-squares regression at each step of the model building stage, it allows rapid automatic exploration of candidate interactions of level two, three or more. In contrast, the same step in MAPLSS is based on PLS rather than OLS. PLS has well-known advantages but it comes with a larger computational cost. For these reasons, we restrict the potential models to those that include interactions no higher than second order. By restricting to this class, MAPLSS can still provide efficient and automatic selection of terms. Certainly this limits the space of possible models, but in practice, with many variables, including all two-way interactions as potential predictors usually suffices. In the same way that PLSS generalizes linear PLS regression, MAPLSS generalizes PLSS regression with the inclusion of all two-way interactions. Including interactions, the design matrix **B** becomes

$$\mathbf{B} = [\underbrace{\ldots \mathbf{B}^i \ldots}_{i \in K_1} | \underbrace{\ldots \mathbf{B}^{k,l} \ldots}_{(k,l) \in K_2}] \tag{6}$$

where K_1 and K_2 are index sets, respectively, for single variables and bivariate interactions. As a result, the fit of the response j can be written as

$$\hat{y}^j(A) = \sum_{i \in K_1} s_i(x^i, \hat{\beta}_i^j(A)) + \sum_{(k,l) \in K_2} s_{k,l}(x^k, x^l, \hat{\beta}_{k,l}^j(A)) \qquad (7)$$

In the case of a small number of predictors, a simple way to visualize the most influential main effects and interactions on the *jth* fitted response is through the inspection of all the possible ANOVA function plots, curves and surfaces, in decreasing order according to the range of all the $s_i(\mathbf{x}^i, \hat{\beta}_i^j(A))$ and $s_{k,l}(\mathbf{x}^k, x^l, \hat{\beta}_{k,l}^j(A))$ transformed data. This presentation is used in MAPLSS to propose a visual interpretation of the nonlinear influence of the retained predictors on the responses. However, to select the most important predictors, an automatic computational approach has been based on both goodness of fit and prediction criteria.

4. THE MODEL-BUILDING STAGE

The initial step of the MAPLSS model building stage consists of building a main effects only model. Eventually, this model will be pruned, but only after two-way interaction terms are considered. Because each basis function is constructed independently, MAPLSS, unlike MARS, CART and other tree-based regression methods, is able to remove main effects during the backward pruning step while retaining the two-way interaction in that same variable.

The PLSS spline parameters (including degree of spline, number and location of knots) are inherited by the MAPLSS model. Starting with the PLSS additive model, the MAPLSS building-model phase constructs the index sets K_1 and K_2 which include the main effects and the bivariate interactions, respectively that enter in the retained model defined in Equations (6) and (7). Because of the number of potential basis functions, the number of candidate models is large, and so one has to use a goodness of fit and prediction criterion with low computational costs. The next section illustrates the main aspects of the computational procedure used by PLS, PLSS and MAPLSS models to decide the best model size as a compromise between fit and complexity.

4.1. Fit and prediction in component-based models

The determination of the model size in the estimation of component-based models is even more difficult in the nonlinear case. In the literature, the usual approach for choosing the best dimension is based on cross-validation. It works by creating G distinct groups of observations $(\mathbf{x}_i, \mathbf{y}_i)$, leaving one group out at a time, and estimating the response of the left out group by creating models on the remaining points. By repeating that procedure until the last group of points has been kept out, the PRESS statistic is computed, that is the sum of the squared residuals of the G models. For MAPLSS, which adds interaction terms one at a time to the purely additive PLSS model, determining the proper dimension to retain using PRESS implies high computational costs. A more direct way of constructing an estimate of the unknown prediction squared errors is to correct average squared residuals (ASR), which, as one example, leads to the Mallows C_p statistics originally proposed as a covariate-selection criterion for the linear regression model. As in MARS, we propose a suitable surrogate to the *PRESS* statistics, using a modified form of the *GCV* criterion (originally proposed by Reference [7]). The GCV statistics are applied here in all three models, linearly with respect to the linear regression on the uncorrelated components $\{\mathbf{t}_k\}_{k=1}^A$. For PLS it is linear with respect to the design matrix $\mathbf{X}$ and in both PLSS and MAPLSS it is linear with respect to $\mathbf{B}$. Compared to the linear model the only thing that changes is that a component $\mathbf{t}_k$ is a linear combination of $\mathbf{B}$ which also depends on the response $\mathbf{Y}$. The *GCV* takes the form

$$GCV(A, \alpha) = \frac{\sum_{j=1}^q \mathrm{ASR}^j(A)}{[1 - \alpha \frac{A}{n}]^2} \qquad (8)$$

where $\mathrm{ASR}^j(A) = \frac{1}{n}\|\mathbf{Y}^j - \hat{\mathbf{Y}}^j(A)\|^2$ is the average squared residuals for the response j modeled with A components and α represents a penalty constant to be fixed. The *GCV* criterion depends on the selected initial space-dimension A as well on α. It is interesting to observe that the approximation $(1-x)^{-2} \sim 1 + 2x$ for small values of $x = \alpha \frac{A}{n}$, leads to the well-known C_p coefficient of Mallows [17] when $\alpha = 1$. In the MARS regression context, Friedman and Silverman [15] report reasons for choosing $2 \leq \alpha \leq 4$. Empirically, in PLSS one has to calibrate α to find values that give $GCV(A, \alpha)$ as close as possible to $PRESS(A)$. The value of α computed in PLSS is typically used in the MAPLSS building-model stage, so that we now denote by $GCV(A)$ instead of $GCV(A, \alpha)$ for simplicity. The challenge remains of finding the usual Occam's razor balance between 'goodness' (of fit and prediction) and 'parsimony' (for both the number A of the retained components and the number of terms entering the model). In order to evaluate the goodness-of-fit, in MAPLSS the well-known $R^2(A)$ criterion, which is the proportion of the total $\mathbf{Y}$ variance accounted for by the components $\mathbf{t}^1, \ldots, \mathbf{t}^A$, does not have anything to add to the $GCV(A)$, being close (or equal) to 1 due to the exponential expansion of the column dimension of the design matrix $\mathbf{B}$. So in order to avoid overfitting problems, we look for parsimonious models with the best values of $GCV(A)$ criterion.

4.2. The forward stage

In the first phase of the MAPLSS building-model stage, we individually evaluate all possible interactions. The criterion for accepting any one candidate interaction is based on the gain in fit and prediction compared to the main effects only model. Then, the selected interactions are ordered in decreasing value and are added one by one to the model only if they improve the GCV of the preceding model by a relative gain of ε. We propose $\varepsilon = 0.2$ based on a simulation study discussed in Subsection 5.1.

Inputs $\varepsilon = 20\%$ the threshold to include or not one interaction; $A_{\max}$ = dimension maximum to explore.

step 0 *Construction of the pure additive PLSS model.* In this preliminary phase only the main effects model is considered. Denoting the main effect model by 'm', decide on the spline parameters as well as on A_m giving the best $GCV_m(A_m)$ value.

step 1 *Individual evaluation of all candidate interactions.* In order to evaluate individually interaction terms, each interaction 'i' is separately added to the main effects. Let '$m+i$' denote the order of the model with one interaction 'i'. Compute $GCV_{m+i}(A)$ and the selection

criterion, CRIT(A_i), as the relative increase of the $GCV_m(A_m)$ after adding an interaction to the main effect model

$$\text{CRIT}(A_i) = \max_{A\in\{1,A_{\max}\}} \frac{GCV_m(A_m) - GCV_{m+i}(A)}{GCV_m(A_m)} \quad (9)$$

Rule: refuse interactions 'i' such that $\text{CRIT}(A_i) < 0$ and order the accepted candidate interactions in decreasing order. Denote $\{i_1, \ldots, i_K\}$ that set, eventually empty, of the accepted interactions ordered by the following inequalities

$$\text{CRIT}(A_{i_1}) \geq \ldots \geq \text{CRIT}(A_{i_K})$$

step 2 *Add successively significant interactions to the pure additive model.* After selecting $\{i_1, \ldots, i_K\}$, the ordered set of candidate interactions, one has to tell whether or not to accept these. Then, the step 2 phase consists in adding to the main effects model, interactions i_1, i_2, etc, successively, provided that one interaction improves the *GCV* criterion of the previous model with respect to the threshold ε.

Finally the following subsection presents the backward phase in order to prune the model of lowest influence ANOVA terms according to the range of the ANOVA functions.

4.3. The backward pruning stage

At the end of the forward phase, different possibilities present themselves for the pruning phase. First, the retained model is purely additive either because all candidate interactions have been refused at step 1, or because of ε in step 2, whose threshold value did not allow inclusion of the largest interaction i_1. Second, the ANOVA decomposition includes some significant bivariate interaction terms and we have to ask the question: do we retain the main effects whose predictors also intervene in some interactions? More generally, the backward phase prunes the model, removing main as well as small effects in the same way as PLSS does, by ordering the ANOVA functions according to the range of the transformations. To obtain the final model (7), an automatic deleting procedure has been used through the selection criterion. However, some users, experts in the scientific domain of the data at hand, prefer manual control of the pruning process in order to better evaluate which interpretable significant ANOVA terms are to be preserved in the final sets K_1 and K_2 characterizing the variables in the model (7).

5. APPLICATIONS AND SIMULATIONS

To illustrate the potential of MAPLSS, we apply it to some simulated and real datasets found in the literature. At first, in order to evaluate the domain of accuracy of the component regression method MAPLSS with respect to the regression tree methods MARS and BRUTO, we use three classical signal functions at three different sample sizes (50, 100, 200). The sample size refers to both the size of the training and test sets. Then we consider a real dataset to show the richness of the MAPLSS output in a multi-response problem and its ability to distinguish important and unimportant predictors.

5.1. Comparison among BRUTO, MARS and MAPLSS on simulated examples

In this section, we present three simulated examples to compare the accuracy of MAPLSS against BRUTO and MARS. We also examine their ability to uncover interaction effects present in the data by looking at three signal functions with no interaction, one interaction and two interactions. The values of predictors in all cases were randomly generated from a uniform distribution and the pure responses were assigned by formula (10, 11, 12), respectively. Because the objective is to compare these methods in a low sample size to variable ratio, we use 100 replications with sample sizes equal to 50, 100, 200 respectively with ten predictors, five of which are used to disturb the signal. Before illustrating the comparison results, we describe the three functions.

The first example ([9], p.34) illustrates what happens when BRUTO, MARS and MAPLSS are applied in situations where the true underlying function is purely additive in the predictor variables (formula 10):

$$f_1(x) = 0.1\exp(4x_1) + \frac{4}{(1+\exp(-20(x_2-0.5)))} + 3x_3 + 2x_4 + x_5 + 0\sum_{i=6}^{10} x_i \quad (10)$$

This function has a nonlinear additive dependence on the first two variable, a linear dependence on the next three and five more predictors are used which do not enter the signal.

The second example from Friedman ([9], p. 37) is used to test the ability of BRUTO, MARS and MAPLSS to uncover interaction effects when they exist.

The model of example 2 (Equation (11)) contains one interaction effect involving the first two predictors. There is a quadratic relationship involving the third predictor, a linear dependence of the fourth and fifth predictors and the last five predictors are independent of the response.

$$f_2(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 1/2)^2 + 10x_4 + 5x_5 + 0\sum_{i=6}^{10} x_i \quad (11)$$

Finally, the third example is a variant of the second, but adds an interaction between x_4 and x_5.

$$f_3(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 20x_4 x_5 + 0\sum_{i=6}^{10} x_i \quad (12)$$

So in all three signal functions, the predictors x_6, x_7, x_8, x_9, x_{10} are noisy. To compare the methods over the simulations, we compute the median and quartiles of the Mean Squared Errors (MSE) of prediction, as a measure of the prediction accuracy, by test sample using observations out of training sample. For each of the signal functions (10, 11, 12), we generate test and training datasets of equal sizes, and compare the MSE distributions via box plots (see Figures 1–3). Because of the large number of tuning parameters, we set the interaction order to 2 for all models. For BRUTO and MARS, we have set the B-spline degree equal to 2 and left all the other parameters at their default values. In MAPLSS, the choice of B-spline parameters has been made by the heuristic strategy (starting with small degree and knot number

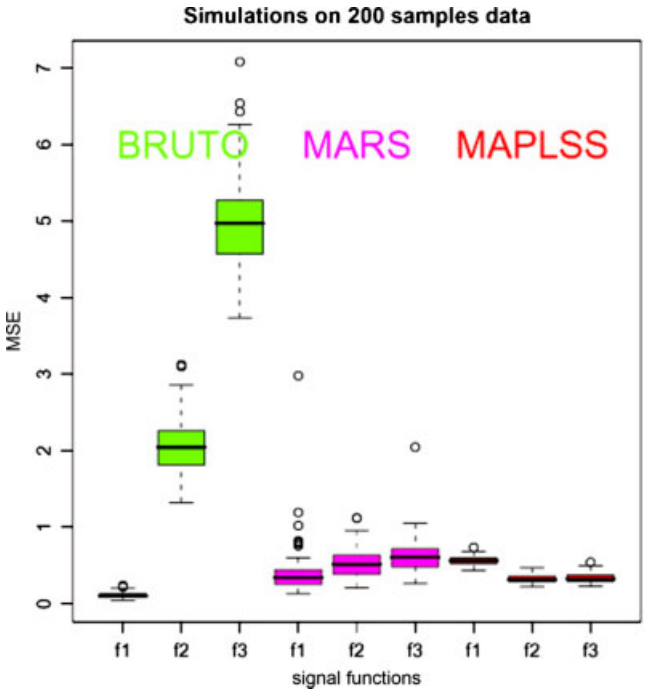

Figure 1. BRUTO, MARS and MAPLSS box-plot distributions of MSE values computed on f1, f2 and f3 signal functions for a 200 sample size.

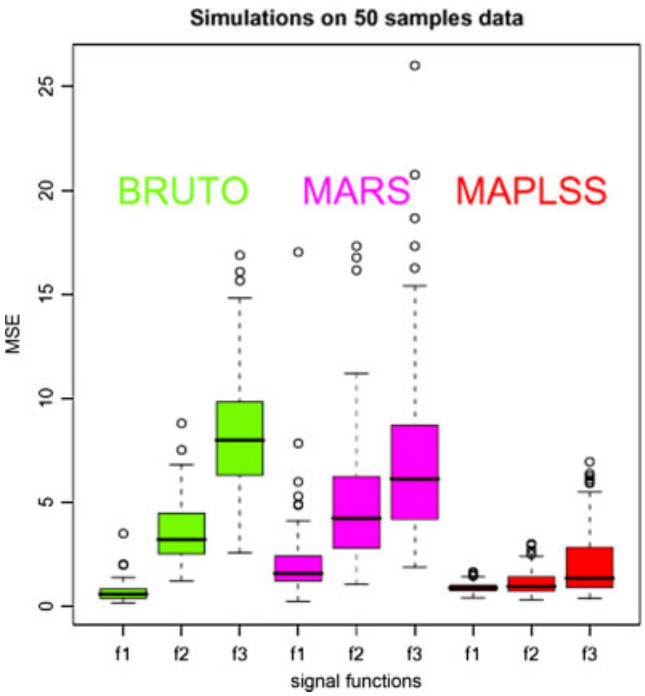

Figure 3. BRUTO, MARS and MAPLSS box plot distributions of MSE values computed on f1, f2 and f3 signal functions for a 50 sample size.

and increasing progressively the model complexity). In all sample cases, the threshold for accepting or rejecting an interaction is 0.2 and the tuning parameter α in *GCV* is equal to 2.

In Figure 1, with sample size 200, MAPLSS performs better than BRUTO and MARS for both f2 and f3 signals while for f1, both BRUTO and MARS models outperform MAPLSS. At sample size 100 (box plots of Figure 2), we note again that with respect to the signals f2 (one interaction) and f3 (two interactions) MAPLSS performs better than MARS and BRUTO, while for the signal f1 (no interaction), BRUTO does best. At the same time for sample size 100 and 200 comparing MARS with BRUTO, we observe that MARS performs better than BRUTO for both f2 and f3 and worse for f1.

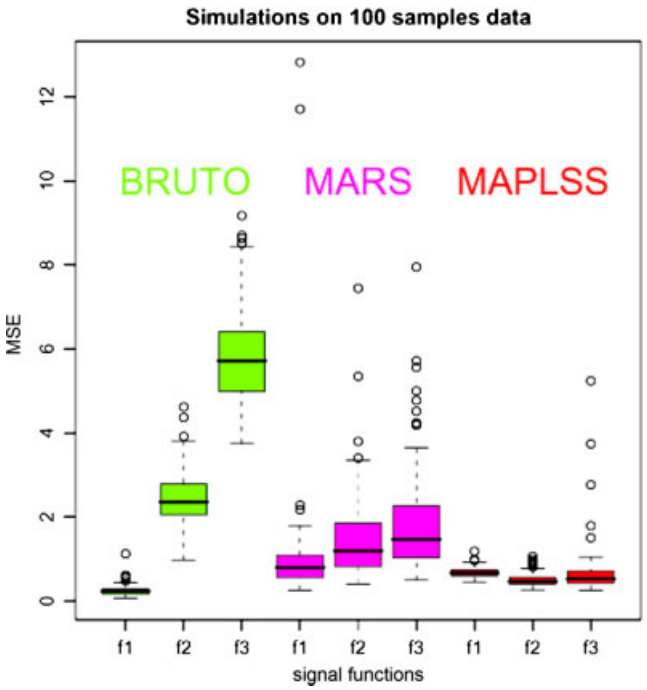

Figure 2. BRUTO, MARS and MAPLSS box-plot distributions of MSE values computed on f1, f2 and f3 signal functions for a 100 sample size.

Finally, the displays of MSE distributions for each signal function, for sample size 50, are given in Figure 3.

We observe that MAPLSS has the best performance as compared to BRUTO and MAPLSS in nearly all cases. While a comparison between MARS and BRUTO shows that MARS performs better than BRUTO for only f3 and worse for both f1 and f2. The box-plots of Figure 3 also point out that the spread of the MSE for MAPLSS model is smaller than both MARS and BRUTO. In conclusion, in all figures (1, 2 and 3), MAPLSS has both a smaller MSE and less variance than both BRUTO and MARS with the exceptions previously mentioned. The box plots show that the BRUTO performances with respect to signals f2 and f3 are the worst for all sample sizes. Also note that when the sample size is increased to 200, (Figure 1) the accuracy of MAPLSS is best when the signal function presents one interaction (f2) or two interactions (f3). The exception is signal f1, where MAPLSS does worse, except in the low sample size case.

5.2. A real example

We now illustrate the capability of MAPLSS on a real data example, in particular on wine evaluation that has ventured sensory-chemical related studies.

The data arise from a 2005 sensory study of a research institute in Campania (Italy) consisting of 15 bottles of white and red wine, involving two sensorial responses to be simultaneously predicted by nine chemical predictors. In this application, chemical and sensory data were collected by an expert panel of judges, to

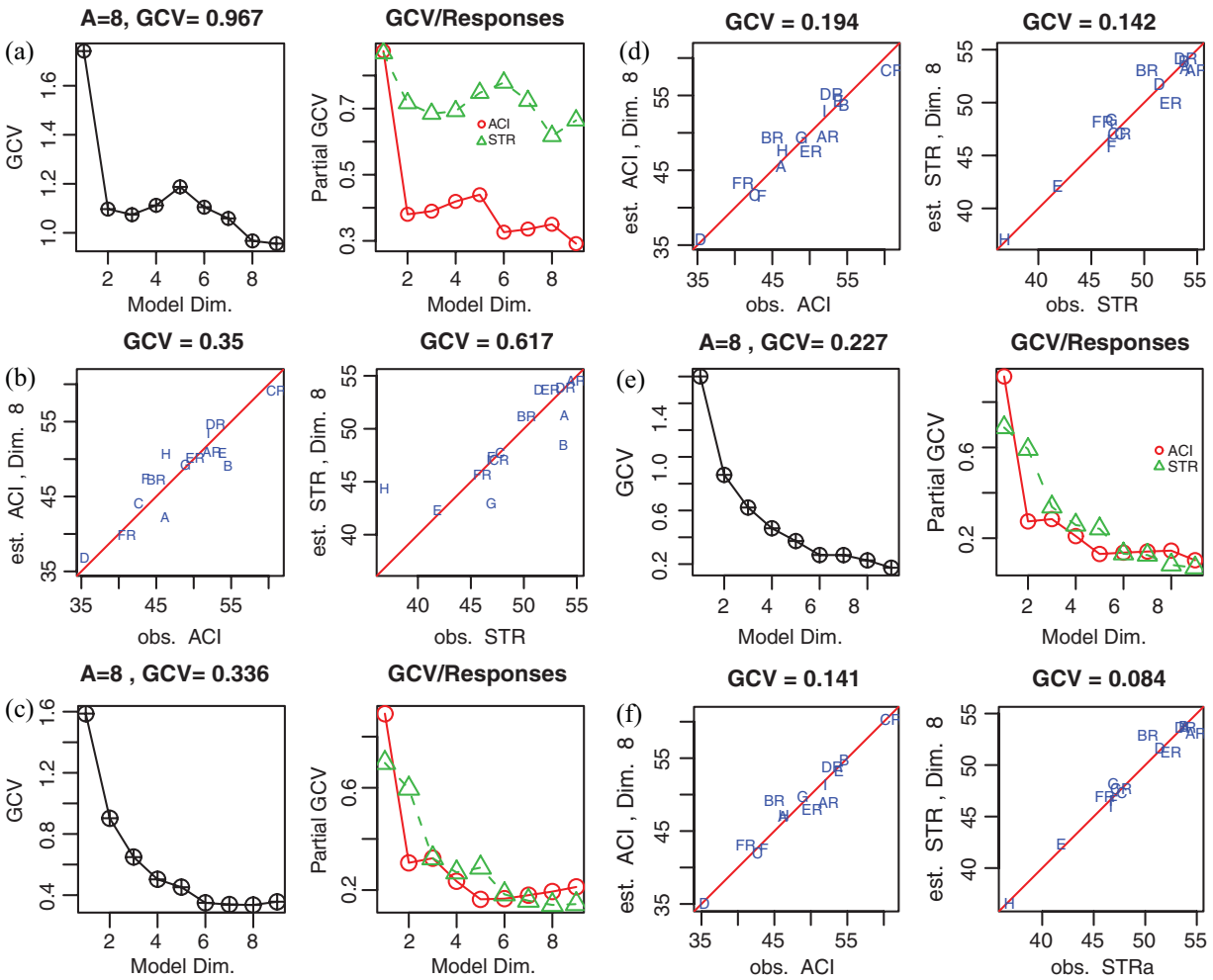

Figure 4. (a) The GCV according to the dimension for the linear PLS multi-response model. Total GCV = 0.967. (b) Observed against predicted values for the two responses in PLS model. Total GCV = 0.967. (c) The GCV according to the dimension for the multi-response PLSS model. Total GCV = 0.336. (d) Observed against predicted values for the two responses in PLSS model. Total GCV = 0.336. (e)The GCV according to the dimension for the multi-response MAPLSS model. Total GCV = 0.227. (f)Observed against predicted values for the two responses in MAPLSS models. After pruning, total GCV = 0.224.

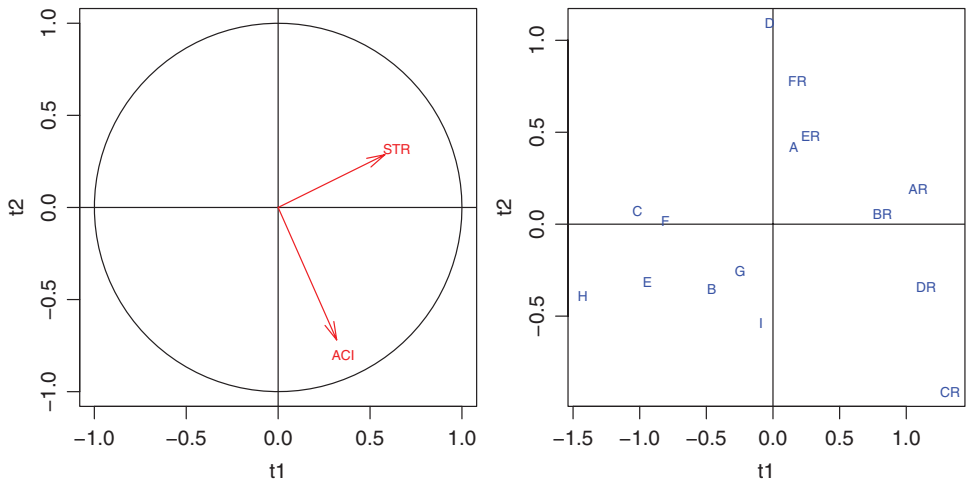

Figure 5. The circle of correlations and the observation plot.

describe and discriminate sensorial stimulus, on commercially available white and red country D.O.C.G. (guaranteed and controlled origin denomination) wines, coming from different areas in Campania. We get nine white wines, **A**,**B**,**C**,**D**,**E**,**F**,**G**,**H**,**I**, and six red wines **AR**,**BR**,**CR**,**DR**,**ER**,**FR**. An expert panel of judges evaluated the following sensorial taste characteristics: acidity **ACI** and structure **STR**. The chemical predictors are the sulphurous anhydride (**SO2**), density (**DEN**), the percentage of alcohol (**ALC**), the dryness (**DRY**), the pH (**PH**), the total acidity (**ACT**) the presence of phenols, in particular flavonoids (**FLA**), polyphenols (**POL**), and proanthocyanidins (**PRO**). All measured variables were scaled to be in the (0, 1) range.

Passing from the linear model, PLS, to the nonlinear ones, without interaction, PLSS, and with interactions, MAPLSS, the accuracy of the final model is greatly improved. In Figure 4a,c,e, we show the GCV prediction plots obtained respectively by PLS,

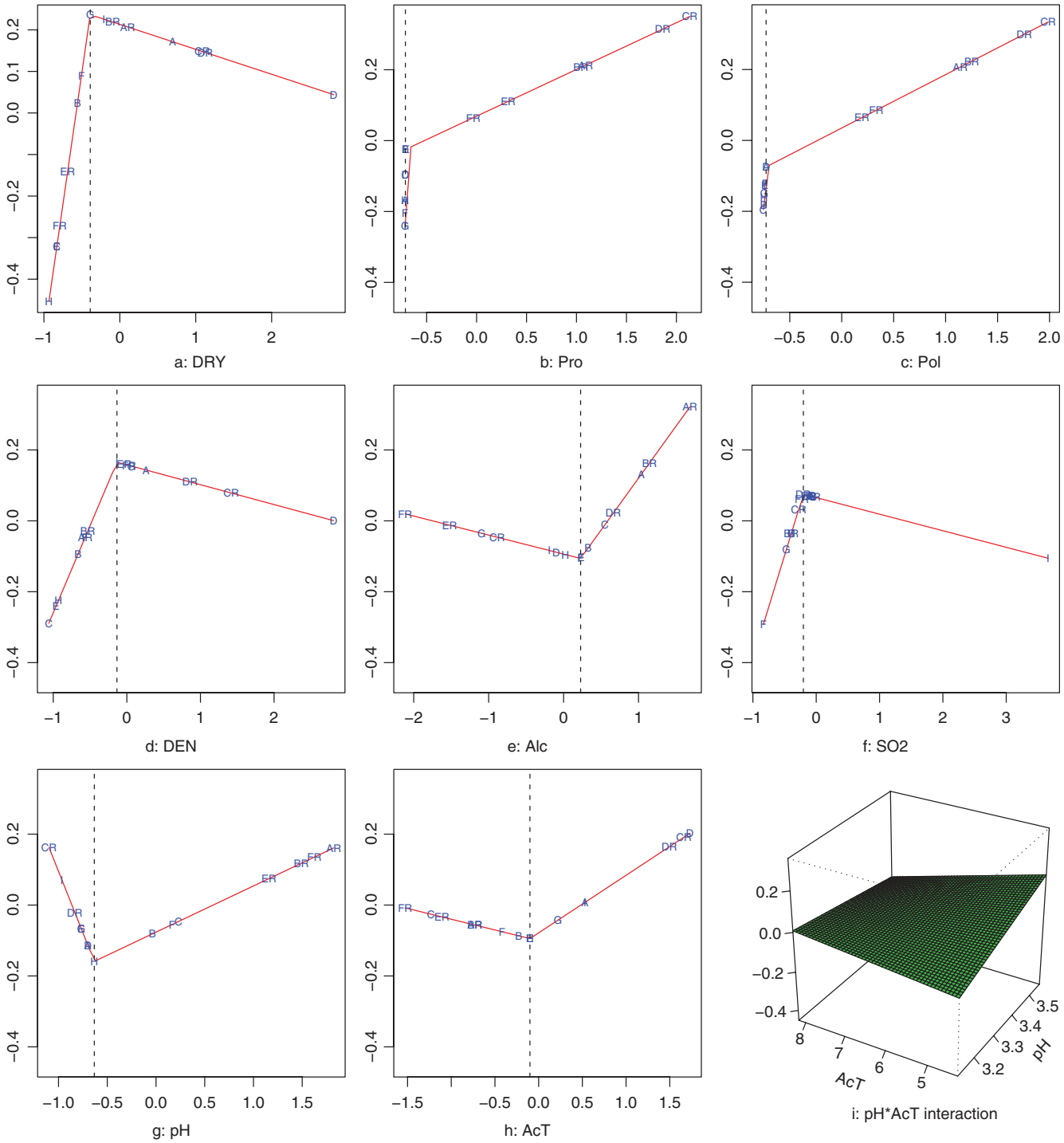

Figure 6. ANOVA plots of the nine retained predictors which more affect the component t1, ordered according to the vertical ranges from left to right and up to down.

Journal of CHEMOMETRICS

PLSS and MAPLSS models and in Figure 4b,d,f we depict the observed against predicted values for each response in PLS, PLSS and MAPLSS, respectively. In particular, observe that the total GCV values range from 96.7% for the linear PLS model to 33.6% for the PLSS main effects, and finally to 22.7% for the MAPLSS model. Concerning the response $y_2 =$ **STR** the prediction is consistently improved in MAPLSS; note that the GCV values range from 61.7% for the linear PLS model (Figure 4b) to 8.4% for the MAPLSS model (Figure 4f). The MAPLSS model was fit using a tuning parameter α in *GCV* equal to 0.5. In MAPLSS, the choice of B-spline parameters is the same as the corresponding main effect PLSS model (phase 0) which follows the heuristic strategy of starting with small degree and knot number and progressively increasing the model complexity. As a result, the degree has been set equal to 1 and the knot number equal to 1, for all predictors. Before including interactions in the PLSS model, the dimension selected was $A = 8$ according to GCV = 0.336. In MAPLSS, after interaction selection (forward and backward phases), we then looked at the predictors

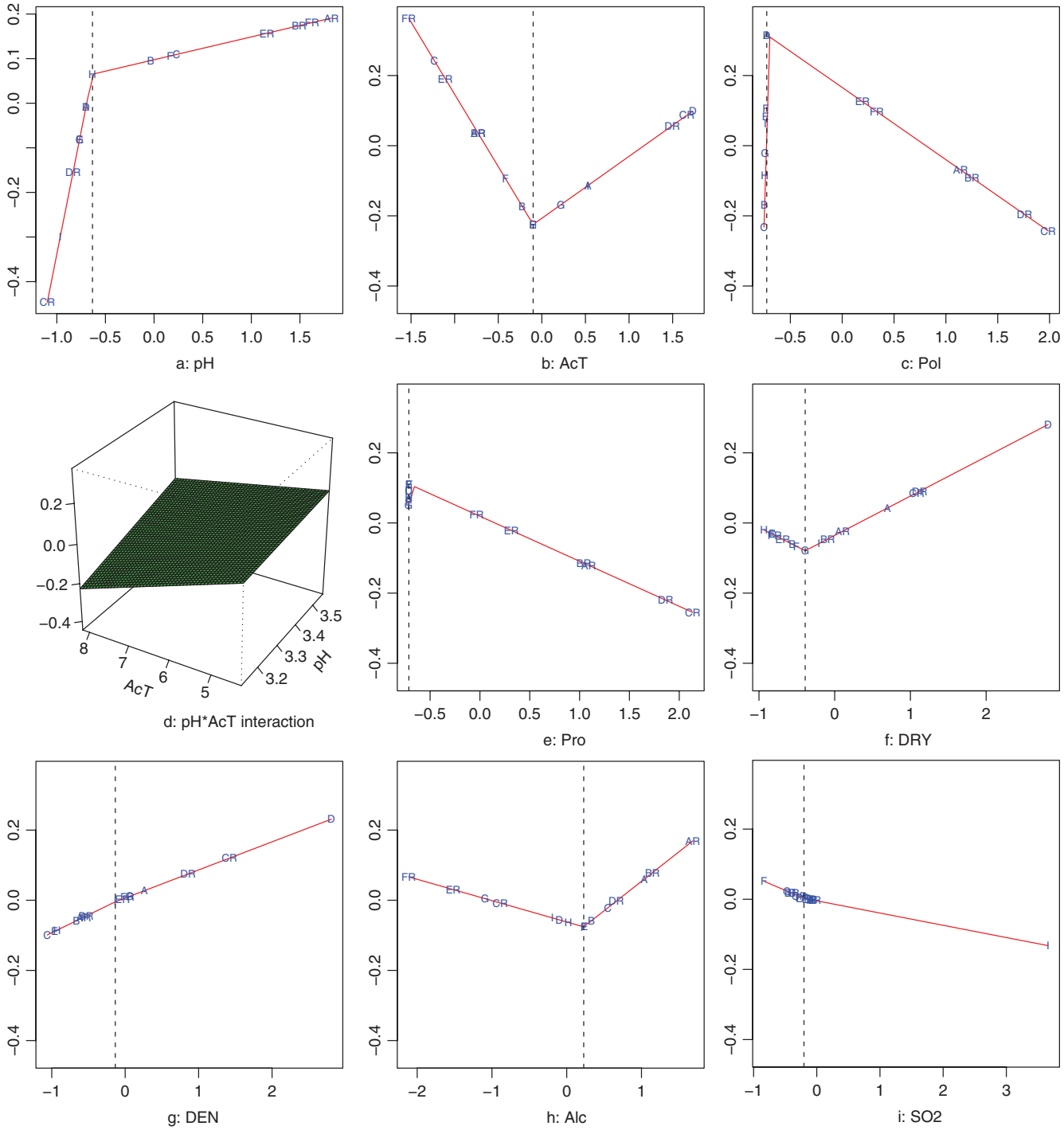

Figure 7. ANOVA plots of the nine retained predictors which more affect the component t2, ordered according to the vertical ranges from left to right and up to down.

which had more influence on the two responses. Among the 36 possible bivariate interactions only one was accepted, **PH** $*$ **ACT**, for all responses. The model choice could also be guided by a subject area expert. As an example, we use the GCV as the criterion for further refinement of the model; before pruning the model we obtain the total criterion $GCV(8) = 0.227$; we decide to prune the model deleting the main predictor **FLA**. After pruning, the GCV criterion does not increase ($GCV(8) = 0.224$), so we decide to retain the simpler model with eight main predictors and one interaction variable. To enrich the interpretation of MAPLSS output, we add some further plots. In Figure 5, we see the circle of correlations and the observation plot of the final MAPLSS model. In the circle of correlations, we note that the two responses are not correlated and this fact helps to explain the model dimension number. Furthermore, the response **STR** is highly correlated with the first component, while the response **ACI** is highly correlated with the second component. Concerning the observation plot, we simply notice two main observation groups in correspondence of

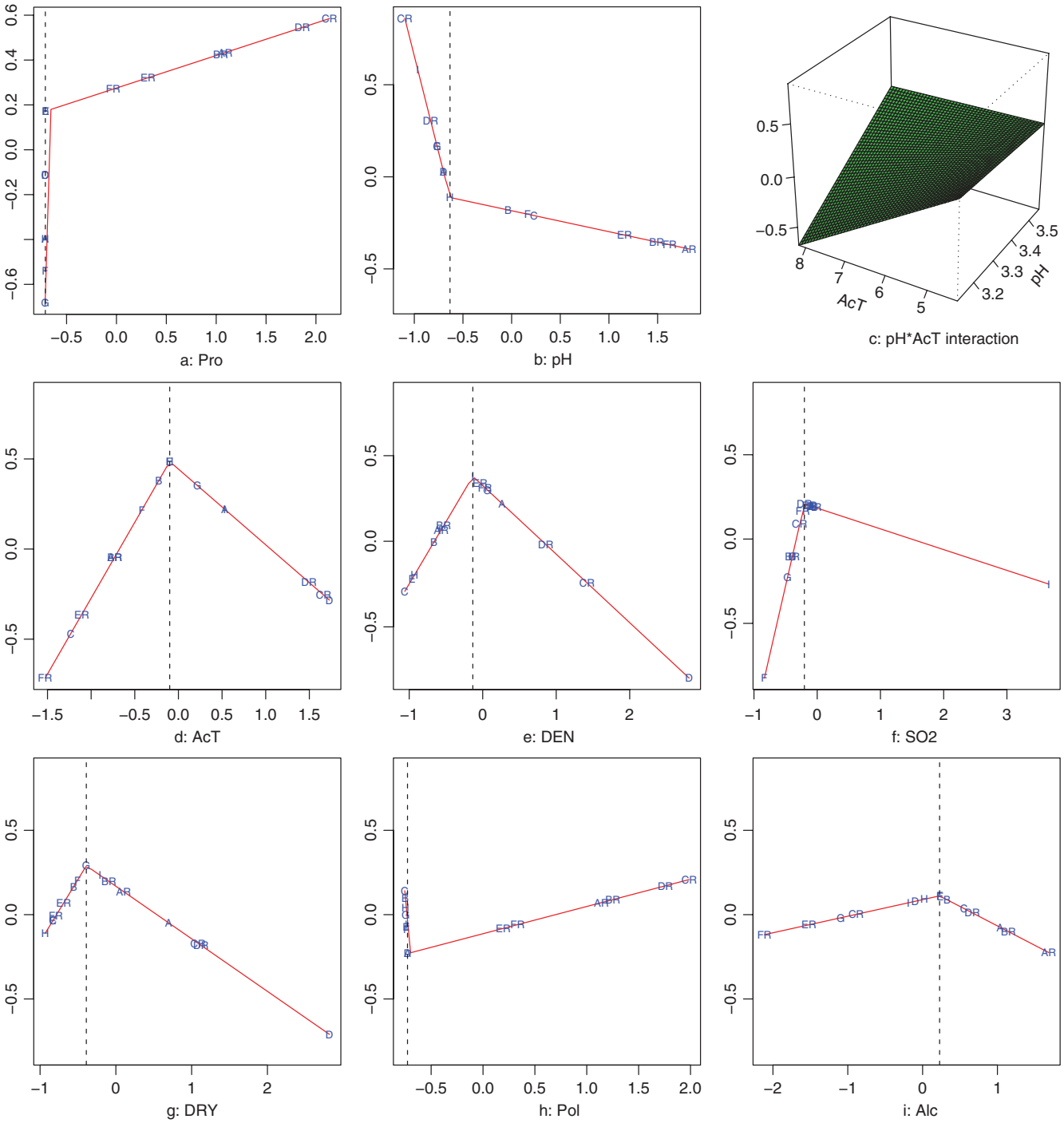

Figure 8. ANOVA plots of the nine retained predictors which more affect the response **ACI**, ordered according to the vertical ranges from left to right and up to down.

white and red wines, and the presence of two white wines (**A**, **D**) in the group of red wines. To better interpret the observation plot, we present the plots of the transformed predictors (Figures 6, 7) with respect to the latent variables **t1** and **t2** (useful to explain the predictor influence on the components or latent variables). Thus in Figures 6 and 7 we give a vision of the nonlinear relationships between the latent variables or components of the model with the predictors. In Figure 6 concerning the influence of predictors on **t1**, the first term in the predictor list is **DRY** followed by **PRO** and **POL** (see Figure 6a,b,c, respectively), etc. We observe that high values of **DRY** given principally not only by red wines but also by two white wines (**A**,**D**) significantly influence the first component **t1**. From Figure 7a to 7f, we look at the effects of predictors on the second component **t2**; this time the first term in the list of the most influential predictors is **PH** followed by **ACT** and by **POL** (see Figure 7a,b,c, respectively), etc. As usual, we

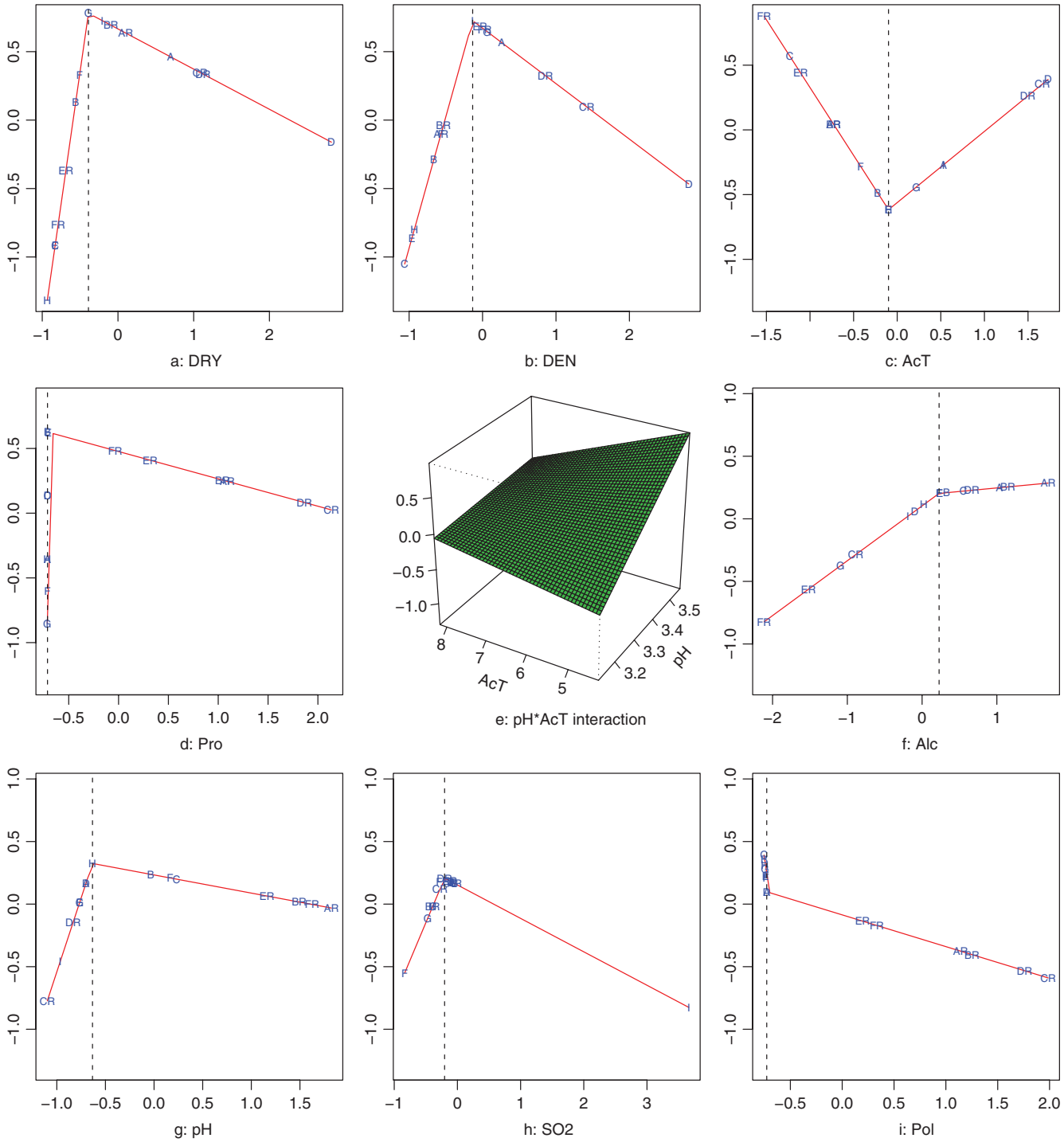

Figure 9. ANOVA plots of the nine retained predictors which more affect the response **STR**, ordered according to the vertical ranges from left to right and up to down.

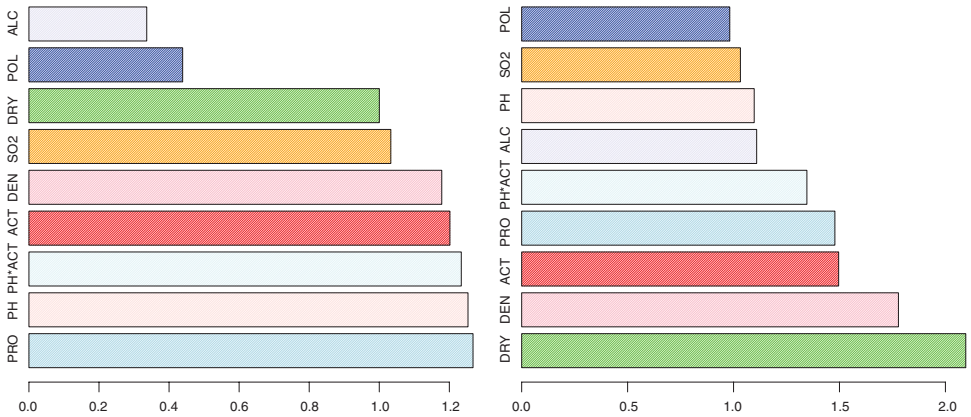

Figure 10. The barplots of the nine retained predictors which affect the responses, **ACI** and **STR**, respectively.

interpret their direct or inverse relationships and their effects on **t2**, on the base of their slopes. For example, large values of **PH** of wines **ER**, **BR**, **FR**, **AR** influence significantly the component **t2**.

In order to understand the influence of the predictors on the responses, in particular with respect to the response **STR** we report the values of the predictor importance before and after pruning the model

1) *Complete Moldel: Main effects plus* **PH** $*$ **ACT**:
GCV = 0.08

DRY	**PRO**	**ALC**	**DEN**	**ACT**	**SO2**	**PH**	**PH*ACT**	**POL**	**FLA**
1.941	1.385	1.033	1.022	0.914	0.891	0.658	0.641	0.542	0.303

2) *Prune low ANOVA terms deleting the main effect* **FLA**:
Final GCV=0.08

DRY	**DEN**	**ACT**	**PRO**	**PH*ACT**	**ALC**	**PH**	**SO2**	**POL**
2.096	1.778	1.496	1.478	1.346	1.109	1.098	1.032	0.982

Notice that to obtain a more parsimonious final model, the main effect **FLA** has been removed despite a non-negligible influence in the first pass model and, as a consequence, in the second pass model, the influence of the interaction, **PH** $*$ **ACT**, is increased.

In Figures 8 and 9, we graphically represent the contributions of the ANOVA terms of the second pass model on the two responses **ACI** and **STR**, respectively. From Figure 8a–f, we look at the effects of predictors on the response **ACI**, interpreting their direct or inverse relationships, and their effects on the base of their slopes. We read for each predictor in the list the low or large values, which belong to different wines, characterizing the responses. For example in Figure 8a, large values of **PRO** of red wines **DR**, **CR** characterize greatly the response **ACI**. Note that the interaction effect, **PH*ACT**, has played a different role in the prediction of **ACI** with respect to **STR**; in fact the third term in the predictor list of **ACI** is the interaction term (see Figure 8c). In Figure 9, the first term in the predictor list is **DRY** followed by **DEN** and **ACT** (see Figure 9a,b,c, respectively),etc. In Figure 9a, low values of **DRY** of white wines **G**, **I** characterize greatly the response **STR**.

At the end, to resume the predictor influence on the responses in Figure 10, we represent graphically the contributions of the ANOVA terms previously shown in Figures 8 and 9, for each of the two responses, using the bar plots of predictor ranges. As you can see, the predictor contributions are different for the pair of responses.

6. CONCLUSION

A statistical method for nonlinear multivariate regression of noisy data in high dimensions, MAPLSS, has been developed in this paper. The MAPLSS offers advantages to current methods especially in cases where predictors are correlated, interaction effects are present, where there is a low sample size to number of variables ratio and where outliers may be present. The performance of the proposed method has been compared to known nonlinear models, as BRUTO and MARS in terms of their accuracy via simulation studies. The multi-responses model, MAPLSS, extends PLSS using an automatic selection of interactions. Like PLS and PLSS, it avoids the problems of both multicollinearity and ill conditioning. Like MARS, it automatically selects the variables and their interaction order. As it has been shown by examples and simulations, MAPLSS is an efficient regression tool in the difficult real-life context.

In our analysis, the MAPLSS model shows the best performance in terms of MSE, for all small datasets (samples of sizes 50, 100, 200) to uncover interactions only when they really exist. At the end, the analysis of the real data has shown that one of the advantages of the multi-response MAPLSS algorithm, being a component regression based method, is that it permits, in a backwards pruning stage, the elimination of predictors, which may unnecessarily complicate the model, but leaving their interactions resulting in a more accurate and more parsimonious model.

7. IMPLEMENTATION

Multivariate Additive Partial Least Squares via Splines has been programmed by the first and second authors in R language. The implementation of the program was easy thanks to the availability of most algorithms which Prof. J. F. Durand developed for PLS via Splines. The real dataset and the program can be downloaded at the web address www.jf-durand-pls.com, or they can be obtained from the first author.

REFERENCES

1. Wold H. Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*, Krishnaiah PR (eds). Academic Press: New York, 1966; 391–420.

2. Tenenhaus M. *La Régression PLS, Théorie et Pratique*, Editions Technip, Paris, 1998.
3. Wold S. Kettaneh-Wold H, Skagerberg B. Non linear partial least squares modeling. *Chemometr. Intell. Lab. Syst.* 1989; **7**: 53–65.
4. Durand JF, Sabatier R. Additive Splines for Partial Least Squares Regression. *J. Am. Stat. Assoc.* 1997; **92**: 440.
5. Durand J.-F. Local Polynomial additive regression through PLS and Splines: PLSS. *In Chemometr. Intell. Lab. Syst.* 2001; **58**: 235–246.
6. Durand JF, Lombardo R. Interaction terms in non-linear PLS via additive spline Transformation. *In Between Data Science and Applied Data Analysis*, Schader P, Gaul J, Vichi M (eds). Springer, 2003; 22–30.
7. Craven P, Wahba G. Smoothing noisy data with spline functions. *Numer. Math.* 1979; **31** 377–403.
8. Hastie TJ. Discussion of Flexible parsimonious smoothing and additive modelling by J. Friedman and B. Silverman. *Technometrics* 1989; **31**: 3–39.
9. Friedman JH. Multivariate adaptive regression splines. *Ann. Stat.* 1991; **19**: 1–141.
10. De Boor C. *A Practical Guide to Splines*. Springer-Verlag: Berlin, 1978.
11. Shumaker LL. *Spline Functions: Basic Theory*. John Wiley & Sons: New York, Chichester, Brisbane, Toronto, 1981.
12. Ramsay JO. Monotone regression splines in action (with discussion). *Stat. Sci.* 1988; **3**: 425–461.
13. Molinari N, Durand JF, Sabatier R. Bounded optimal knots for regression splines. *Comput. Stat. Data Anal.* 2004; **45**(2): 159–178.
14. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning, Data Mining, Inference and Prediction*, New York, Springer series in Statistics, 2001.
15. Friedman JH, Silverman BW. Flexible parsimonious smoothing and additive modeling. *Technometrics* 1989; **31**: 3–39.
16. Gifi A. *Non-linear Multivariate Analysis*. Wiley: Chichester, 1990.
17. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. Chapman & Hall/CRC: New York, 1990.
18. Stone CJ. Additive Regression and other nonparametric models. *Ann. Stat.* 1985; **13**: 689–705.
19. Eubank RL. *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York and Basel, 1988.