Bootstrap selection of Multivariate Additive PLS Spline models

Jean-François Durand
Montpellier II University, France
E-Mail: support@jf-durand-pls.com

Abdelaziz Faraj
Institut Français du Pétrole
E-Mail: abdelaziz.faraj@ifp.fr

Rosaria Lombardo
Second University of Naples, Italy.
E-Mail: rosaria.lombardo@unina2.it

The context :

Expensive cost of data points that are oil spots used to forecast crude oil production.

The aim is to build

- an effective nonlinear statistical model

based on

- an as small as possible training data set.

The tools :

- boosted PLS models,

constructed with an

- incremental selection of training samples.

Boosted PLS regression [2, J.F. Durand]

What is L_2 Boosting? [5, Hastie et al.]

The training sample: $\{y_i, \underline{x}_i\}_1^n$,

$$y \in I\!R, \quad \underline{x} = (x^1, \ldots, x^p) \in I\!R^p, \quad \text{centered},$$

to estimate F^* restricted to be of "additive" type,

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

that minimizes the expected L_2 cost

$$I\!E[C(y, F(\underline{x}))] \quad C(y, F) = (y - F)^2/2\ .$$

- M is the "dimension" of the additive model.
  $\mapsto$ cross-validation (CV) or generalized cross-validation (GCV)
- algorithm: repeated (M-1 times) least-squares fitting of residuals (the pseudo-responses).
- the base learner $h(\underline{x}, \underline{\theta})$, a parametrical function of $\underline{x}$ (generally used to capture nonlinearities and interactions).

The PLS base learners: latent variables

$\mapsto$ Usual Partial Least-Squares (PLS), [10, Wold et al.]

PLS belongs to the L_2-boosted methods, by the choice

$$\underline{\theta} \in I\!R^p \ , \ t = h(\underline{x}, \underline{\theta}) =< \underline{\theta}, \underline{x} >= \sum_{j=1}^{p} \theta^j x^j \ ,$$

where, $cov(t, y)$ is maximum.

The linear fit depending on the dimension M

$$\hat{y}(M) = F(\underline{x}\,, M) = \sum_{j=1}^{p} \beta^j(M) x^j.$$

Denoting $Y \in I\!R^{n\times 1}$, $X \in I\!R^{n\times p}$ the training data matrices, the PLS regression is summarized as

$$PLS(X, Y).$$

$\mapsto$ PLS specificity: $M-1$ successive deflations of X, t is a linear compromise of pseudo-predictors.

$\mapsto$ Partial Least-Squares Splines(PLSS), [1, JF Durand]

A predictor x^j is transformed by a B-splines family

$$\{B_1^j(x^j), \ldots, B_{r_j}^j(x^j)\}.$$

Denote B^j the $n \times r_j$ centered coding matrix of the x^j's sample.

$$B = [B^1 | \ldots | B^p]$$

is the super-coding matrix of the predictors. Then

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

The PLSS base learner becomes

$$t = h(\underline{x}, \underline{\theta}) = \sum_{j=1}^{p} \sum_{k=1}^{r_j} \theta_k^j B_k^j(x^j) = \sum_{j=1}^{p} h_j(x^j)\,.$$

The PLSS fit capturing the main effects additively

$$\widehat{y}(M) = \sum_{j=1}^{p} \sum_{k=1}^{r_j} \beta_k^j(M) B_k^j(x^j) = \sum_{j=1}^{p} s_M^j(x^j),$$

$s_M^j(x^j)$ is the 'coordinate' spline function of the influence of x^j on y.

$\mapsto$ Multivariate Additive PLS Splines (MAPLSS) [3, Durand & Lombardo][6, Lombardo, Durand, De Veaux]

To capture bivariate interactions, MAPLSS incorporate two by two tensor products of B-splines families.

The centered main effects + interactions coding matrix:

$$B = [B^1|\ldots|B^p\,\|\ldots|B^{j,j'}|\ldots]$$

where (j, j') belongs to the set $\mathcal{I}$ of accepted couples of interactions

$$MAPLSS(X, Y) \equiv PLS(B, Y)$$

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

A latent variable t becomes a sum of univariate and bivariate spline functions

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

The MAPLSS model is casted in the ANOVA type decomposition

$$\hat{y}(M) = \underbrace{\sum_{j=1}^{p} s_M^j(x^j)}_{\text{main effects}} + \underbrace{\sum_{(j,j')\in\mathcal{I}} s_M^{j,j'}(x^j, x^{j'})}_{\text{bivariate interactions}} .$$

The building model stage incorporates

- an automatic detection of $\mathcal{I}$ (relevant interactions)
- the selection of M by using CV (or GCV) criterion.

Incremental selection of training sets

Paradigm

The principle of the iterative design of experiments is based on the maximisation of the variance of prediction estimated on a set of points designated as candidates.

This criterion finds its justification in the algorithms of exchange used in the computation of optimal designs of experiments ([7, Fedorov]; [8, Mitchell]). For example, in [4, Gilardi and Faraj], the variance of prediction is calculated by a committee of neuron networks.

Iterative algorithm

Step 0: A small training sample and a set of external candidates
REPEAT
Current Step

- Construct m MAPLSS models on bootstrapped training data, and compute for each candidate the m predicted values.
- Incorporate the candidates of largest variance.

UNTIL the sequence of PRESS values stabilizes.

The reservoir simulator data

The results of the adaptive design of experiments are shown on the reservoir simulator data [9, Scheidt et al] based on 10 predictors to forecast oil production.

• Starting with 12 training data (step 0) that define the range $[-1, +1]$ of the predictors,

• The current step implements $m = 10$ bootstrapped MAPLSS models

• 5 new data are added at the end.

The decision to stop for the final training set is based on the stabilization of the PRESS (step 5, 37 observations).

Figure 1 displays the boxplots of 11 Predictive Error Sum of Squares (PRESS) values computed, at each iteration, on the training set and on $m = 10$ bootstrapped drawings.

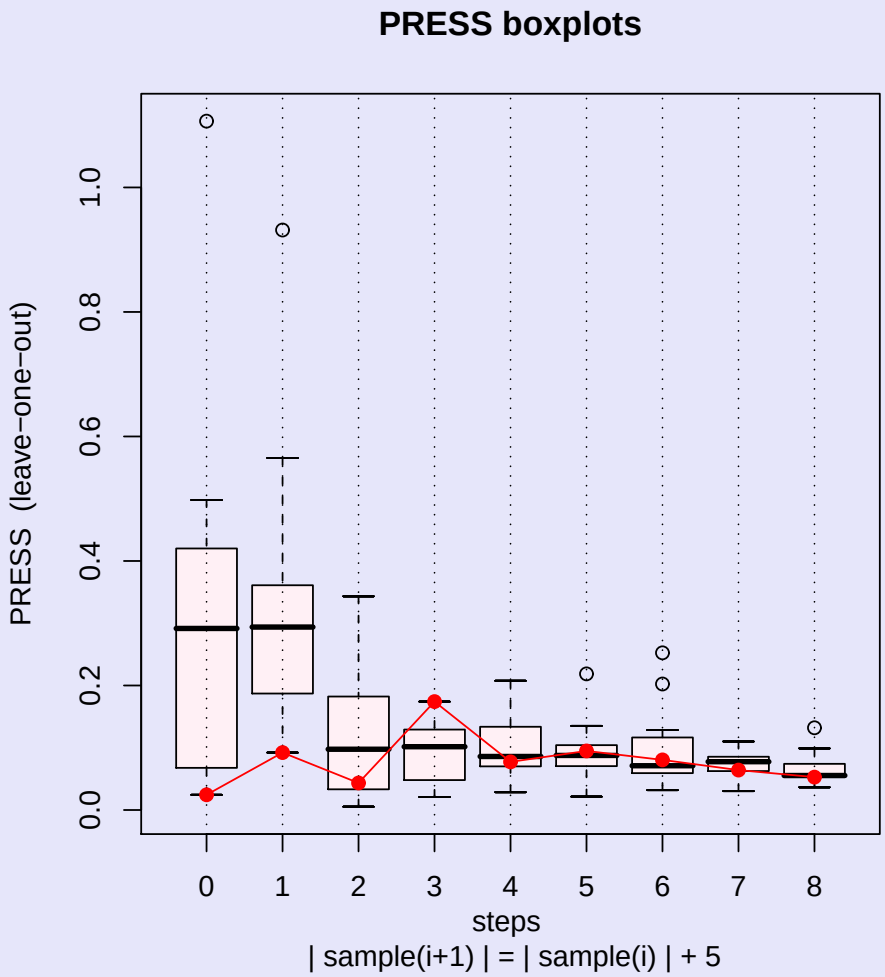
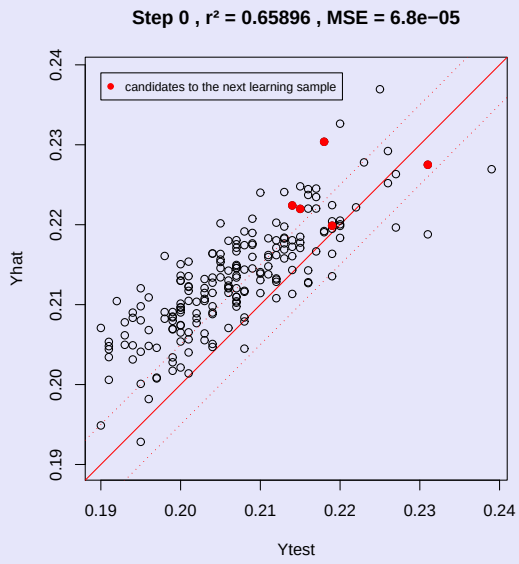
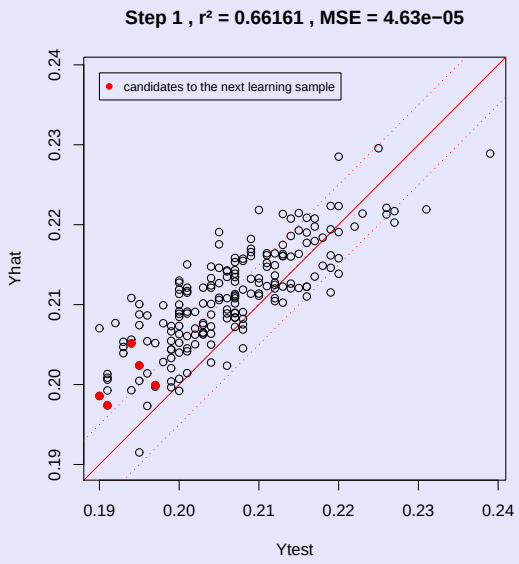
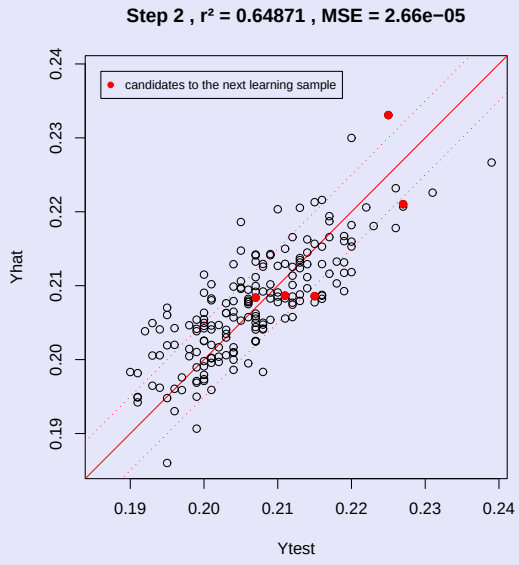
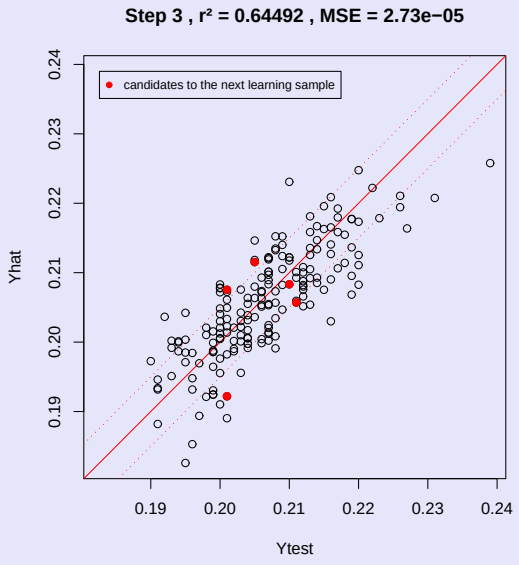


Figure 1: Boxplots of the PRESS values computed, at each iteration, on the training set (red point) and on its bootstrapped drawings.


Step 0 , r² = 0.65896 , MSE = 6.8e−05
candidates to the next learning sample
Yhat
Ytest
0.19
0.20
0.21
0.22
0.23
0.24


Step 1 , r² = 0.66161 , MSE = 4.63e−05
candidates to the next learning sample
Yhat
Ytest
0.19
0.20
0.21
0.22
0.23
0.24


Step 2 , r² = 0.64871 , MSE = 2.66e−05
candidates to the next learning sample
Yhat
Ytest
0.19
0.20
0.21
0.22
0.23
0.24


Step 3 , r² = 0.64492 , MSE = 2.73e−05
candidates to the next learning sample
Yhat
Ytest
0.19
0.20
0.21
0.22
0.23
0.24

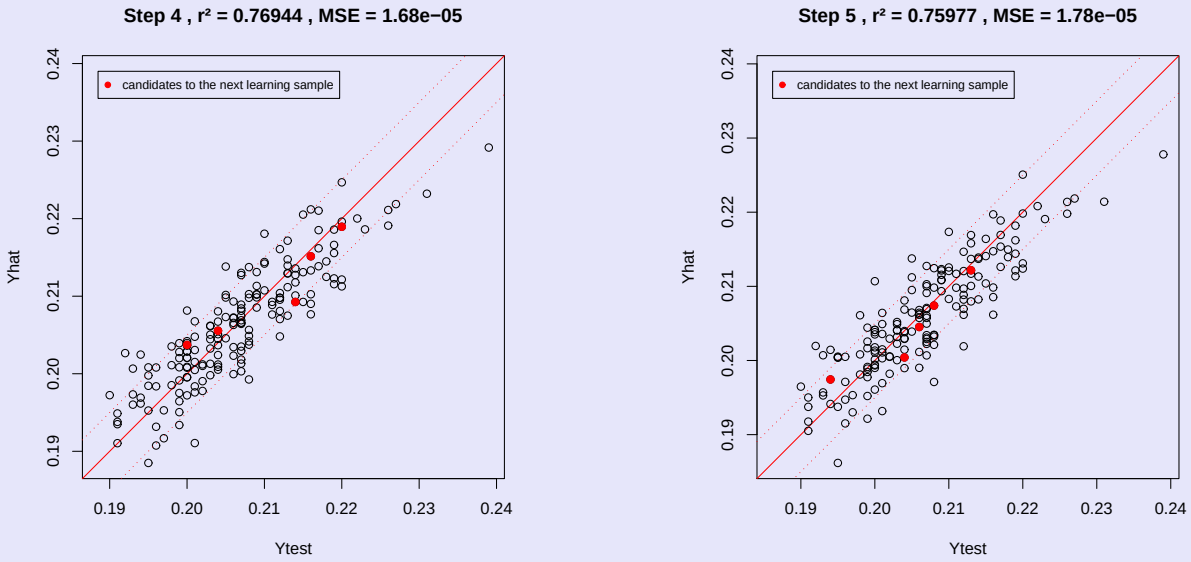


Figure 2: $(Yhat, Y)$ plot of test data at each step. The five red points mark the candidates that will be accepted in the next training set.

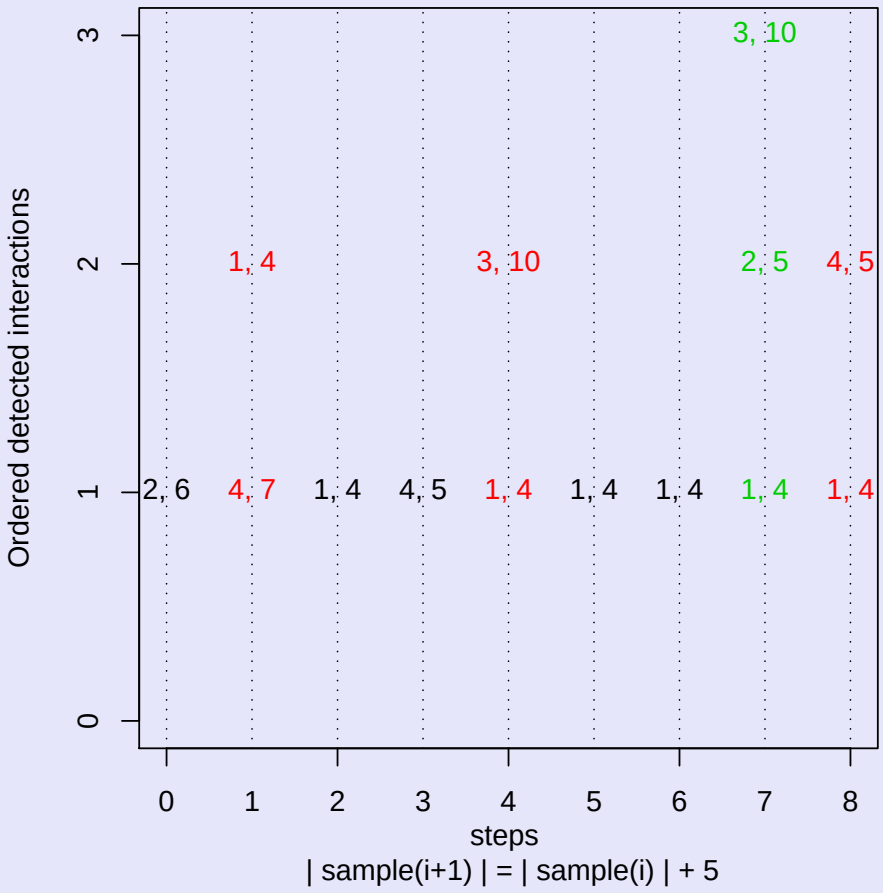


Figure 3: Ordered interactions detected by MAPLSS at each step, interaction $(1, 4)$ is the most relevant.

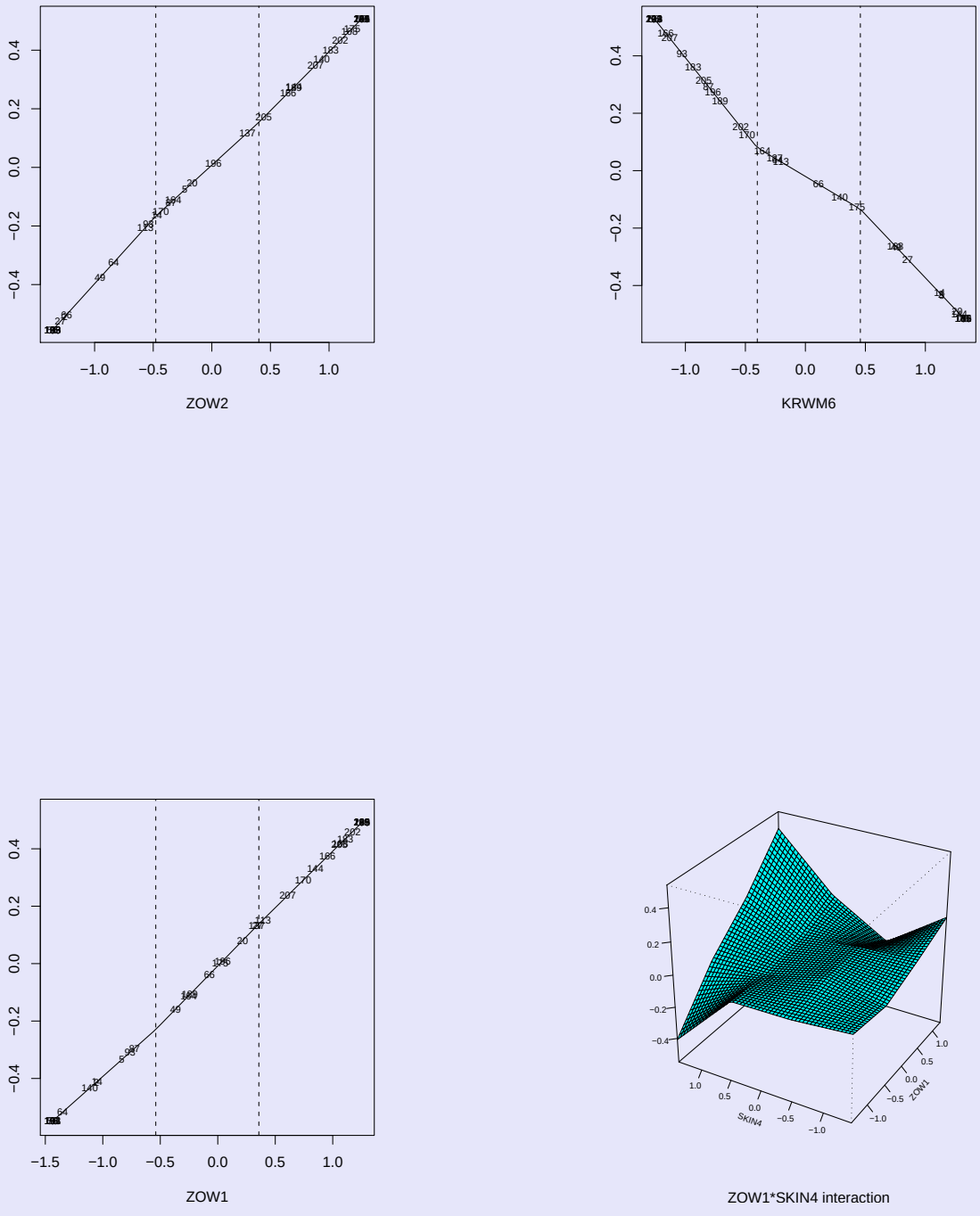


Figure 4: Retained MAPLSS model (step 5): first four main effects and interactions plots, decreasingly ordered from left to right and up to down.

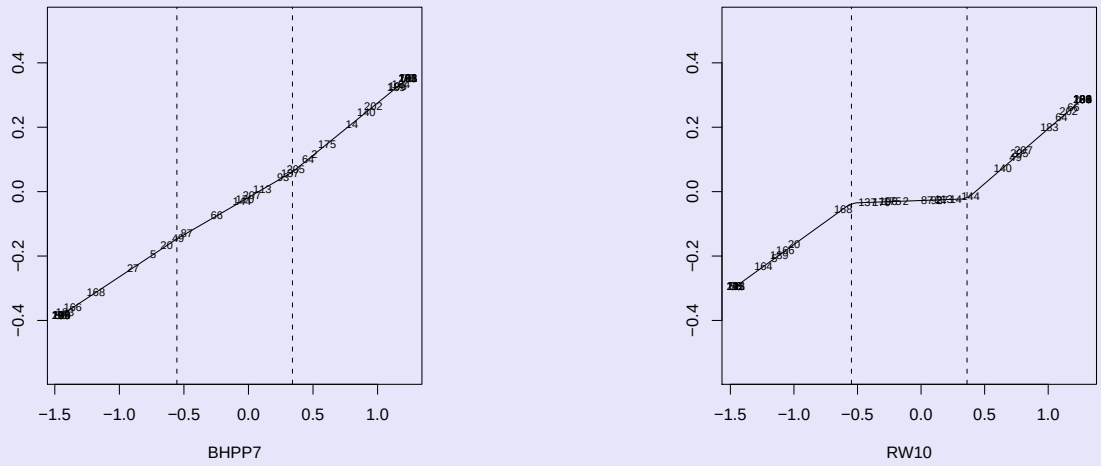


Figure 5: Retained MAPLSS model (step 5): plots of main effects number 5 and 6.

References

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

[2] J. F. Durand (2008), La régression Partial Least-Squares boostée, *Revue MODULAD*, 38, 63-86.

[3] J. F. Durand, R. Lombardo (2003), Interactions terms in nonlinear PLS via additive spline transformations. *In "Studies in Classification, Data Analysis, and Knowledge Organization", Springer-Verlag*, 22-29.

[4] N. Gilardi, A. Faraj (2004), Design of experiments by committee of neural networks, *IEEE International Joint Conference on Neural Networks*, Budapest, 25-29.

[5] T. Hastie, R. Tibshirani, J.H. Friedman (2001), *The Elements of Statistical Learning*, Springer.

[6] R. Lombardo, J. F. Durand, D. De Veaux (submitted), Multivariate Additive Partial Least-Squares Splines, MAPLSS.

[7] V.V. Fedorov (1972), *Theory of Optimal Experiments*, Academic Press, New-York.

[8] T.J. Mitchell (1974), An algorithm for the construction of D-optimal experimental designs, *Technometrics*, 16, n2, 203-210.

[9] C. Scheidt, I. Zabalza-Mezghani, M. Feraille, B. Guard, D. Collombier (2006), Adaptive experimental design for non-linear modeling. *Application to quantification of risk for real field production, Proceedings ECMOR X Amsterdam*, 4-7.

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