Bootstrap selection of Multivariate Additive PLS Spline models

Jean-François Durand
Université Montpellier II
jf.durand@club-internet.fr

Abdelaziz Faraj
Institut Français du Pétrole
abdelaziz.faraj@ifp.fr

Rosaria Lombardo
Second University of Naples
rosaria.lombardo@unina2.it

Abstract: *Multivariate Additive PLS Splines, in short MAPLSS, are Partial Least-Squares models that study the dependence of a set of responses on spline transformations of the predictor variables which permit to capture additively non linear main effects and interactions. The aim of this paper is to present a way of selecting MAPLSS models through an adaptive incremental selection of training samples by a bootstrap procedure. This approach is attractive in the case of expensive data thus implying to construct efficient models based on small training data sets.*

Keywords: Bootstrap, PLS regression, B-splines, Design of experiments.

1. Iterative design of experiments with MAPLSS

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. Those points with a high variance are considered as being potentially informative because they point out insufficiently representative – or insufficiently knowledgeable – regions in the input domain for the model. In the new training sets, we consider the points whose predicted values are of maximum variance, because the variance of the predicted values of new experiments becomes smaller.

This criterion finds its justification in the algorithms of exchange used in the computation of optimal designs of experiments (Fedorov, 1972; Mitchell, 1974). For example, in Gilardi and Faraj (2007) the variance of prediction is calculated by a committee of neuron networks.

The crucial reason to enhance this criterion is the necessity of building effective statistical models based on small training sets due to the expensive cost of data points. In that context, the Partial Least-Squares (PLS) regression of Wold (Wold et al., 1983) is well known to provide robust linear models when the observations/variables ratio is small.

The aim of this paper is to show how an iterative design of experiments works when applied to a recent extension of PLS to non linearity by the use of B-splines, (Durand, 2001, 2008; Durand & Lombardo, 2003; Lombardo et al., submitted). This extension called Multivariate Additive PLS Splines, in short MAPLSS, provides non linear additive models to capture main effects and relevant interactions. Using M components, the fit of the jth response is cast in the ANOVA type decomposition

$$\hat{y}_M^j = \sum_{i=1}^{p} s_M^{j,i}\left(x^i\right) + \sum_{(i,i')\in I} s_M^{j,ii'}\left(x^i, x^{i'}\right)$$

based on univariate and bivariate spline transformations of the predictors, the relevant couples of interactions are selected in an automatic way.

Here, in the one-response case, a bootstrap procedure is used on a given training set to select new observations as candidates to belong to the training sample. Given, say, *m* models based on bootstrapped data, the key point is to add some new observations whose *m* predicted response values are of largest variance.

2. The reservoir simulator data

The results of the adaptive design of experiments are shown on the reservoir simulator data (Scheidt et al., 2006) based on 10 predictors to forecast oil production. Figure 1 displays the boxplots of *m=11* Predictive Error Sum of Squares (PRESS) values computed, at each iteration, on the training set and on 10 bootstrapped drawings. Starting with 12 training data (step 0), 5 new data are added at each step. The decision to stop for the final training set is based on the stabilization of the PRESS (step 5, 37 observations).

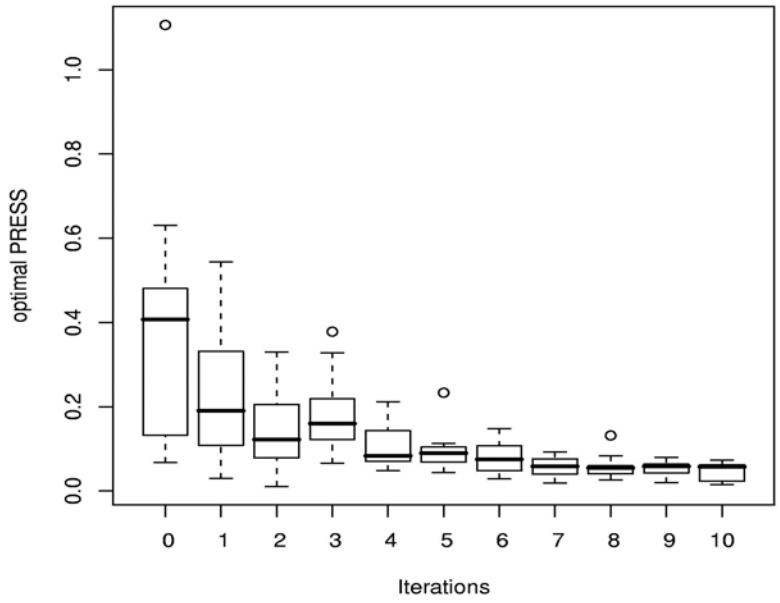

Figure 1: The reservoir simulator data: Boxplots of the PRESS values computed, at each iteration, for the different MAPLSS models based on bootstrapped data. Step 0, the training set is made of 12 samples that define the range [-1,+1] of the predictors. At each step, 5 new observations of largest prediction variance are added.

Bibliography

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

Durand J.F. (2008), La régression Partial Least-Squares boostée, Revue MODULAD, 38, in the press.

Durand J.F. & Lombardo R. (2003), Interaction terms in non linear PLS via additive spline transformations, in: Studies in Classification, Data Analysis and Knowledge Organisation, Between Data Science and Applied Data Analysis, Schader M., Gaul W. & Vichi M. (Eds.), Springer, 22-29.

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

Gilardi N. & Faraj A. (2004), Design of experiments by committee of neural networks, IEEE International Joint Conference on Neural Networks. Budapest 25-29 July 2004, Hungary.

Lombardo R., Durand J.F. & De Veaux R.D., Multivariate Additive Partial Least-Squares Splines, MAPLSS, submitted.

Mitchell T.J., (1974), An algorithm for the construction of D-optimal experimental designs, Technometrics, **16**, n°2, 203-210.

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

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

