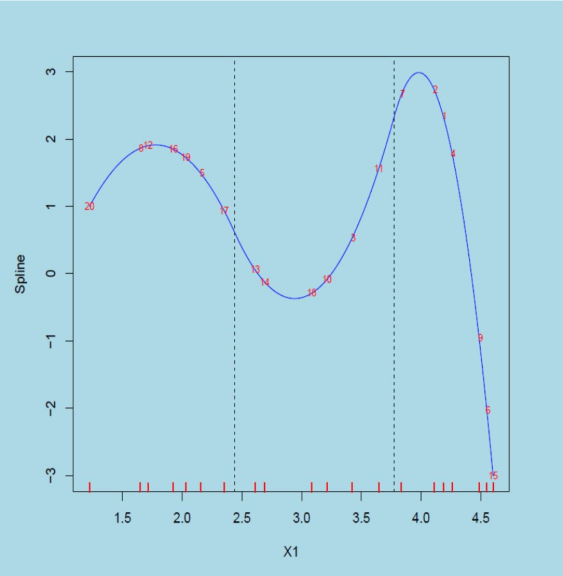
Boosted PLS regression for Prediction and Data Analysis

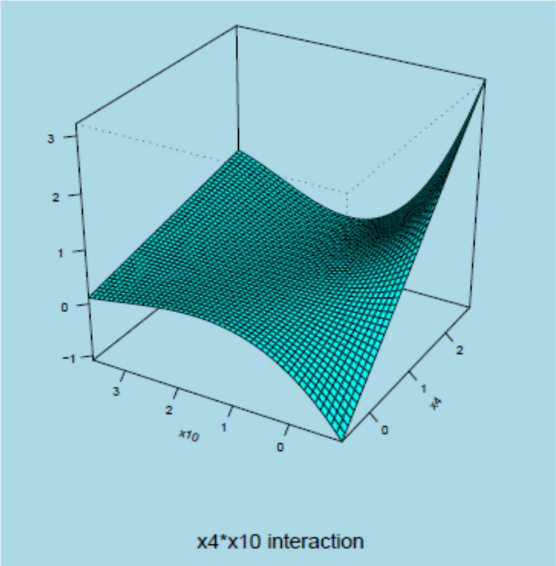
www.jf-durand-pls.com
jf.durand001@orange.fr

Developed at the Montpellier 2 university, the PLSS package (PLSS for Partial Least-Squares Splines) allows to construct statistical models of prediction whose domain of interest concerns many scientific and industrial areas such as chemistry, near infrared spectroscopy, sensometrics, econometrics, marketing, credit scoring.

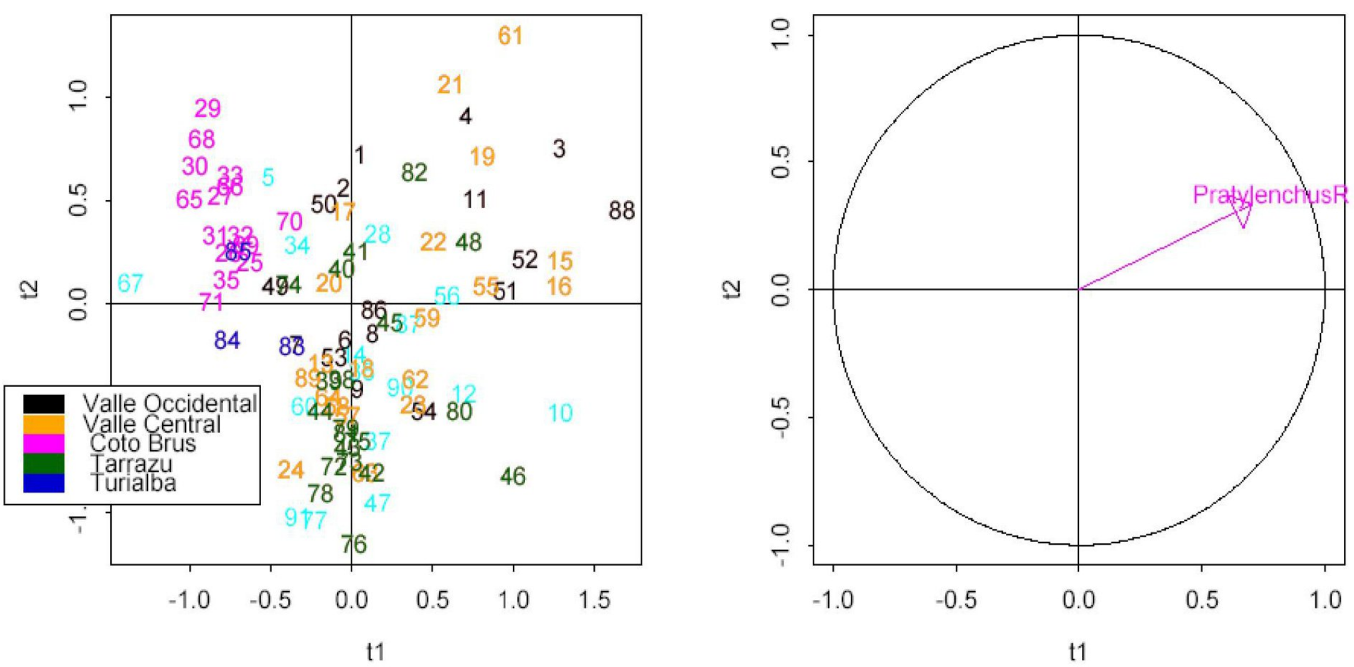
Inheriting good properties from the Partial Least-Squares Linear models (PLSL) elaborated by H. and S. Wold during the eighties, PLSS differentiates itself from other methods as an efficient prediction tool in the following difficult statistical context : a set of predictors characterized by their large number, sometimes several hundred, and also by their heterogeneous nature (a mixture of continuous, categorical and binary variables), all measured on few observations used to predict one or several responses that can be continuous, leading in that case to regression models, or categorical, leading to decision models.

PLSS is new in the sense that transforming the predictors with spline functions (piecewise polynomials), allows the models to capture linear or nonlinear relationships between the responses and the most influential predictors called the main effects variables and their possible interactions.





Besides its characteristics of a tool for prediction, PLSS proposes to look at the data through 2-D maps called the "component scatterplots" and built from principal components.



Principal components also called latent variables or base learners, used to predict the responses, are few independent synthetic variables that are the sum of the linearly or non-linearly transformed predictors and their eventual relevant interactions.

The user is able to experiment more desirable new responses through the use of local or global, continuous or discontinuous, on-line transformations of the observed response based on some variations around the identity spline function (for more details, see the “Short guide to the function *Bsplines*”).

The timing of the prediction process can be split up into 2 steps:

1. Set up the aims of the problem and the associated schedule conditions.
2. The building-model phase: an (a)-(b)-(c) round-trip until obtaining a training data set leading to validated models.
    a) Building an evolutionary data base following the retained schedule conditions.

b) Data processing by the Boosted PLS package and validation or not of the models built on the data at hand.

c) Elaborate some scenarios of prediction. A scenario allows the user to enter new real or fictive data (including eventual different response values close to the observed ones) and test the validated models.

Some industry/research contracts :

2008 - Institut Français du Pétrole: Forecasting oil production by using an adaptive design of experiments.

2008 - Institut Français de la vigne et du vin : Elaboration d'un modèle de maturité du cépage Mourvèdre.

2005 - Hospital of Aversa (Italy): Evaluation of patient satisfaction in health services.

2004 - Oeneo group (France): Identifying the main influential processing factors on the permeability and the mechanical properties of the Altec cork.

2003 - INRA-CIRAD : Modeling the parasitic attacks of the coffee tree roots in Costa-Rica.

2002 - ITV France : Models of the vintage quality in the Bordeaux area.

2000 - DANONE group - TEPRAL research center : Sensorial analyses of orange juices.

Short Bibliography :

Durand, J.F., Lombardo R., Camminatiello, I., (2025), "Identity spline variations in boosted Partial Least-Squares: a study on poverty", Statistical Methods & Applications, 34, 1077--1093.

Durand, J.F., (2001), "Local Polynomial Additive Regression through PLS and Splines: PLSS", Chemometrics and Intelligent Laboratory Systems, vol. 58, issue 2, 235-246.

Durand, J.F., (2002), "Eléments de calcul matriciel et d'Analyse Factorielle de Données", cours polycopié, Université Montpellier II, France.

Durand, J.F., (2008), "La régression Partial Least-Squares boostée", revue MODULAD, 38, 63-86.

Durand, J.F. and Lombardo, R., (2003), "Interactions terms in nonlinear PLS via additive spline transformations", Between Data Science and Applied Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization. Eds M. Schader, W. Gaul and M. Vichi, Springer, 22-29.

Durand, J.F. and Sabatier R. (1997), "Additive splines for partial least squares regression". Journal of the American Statistical Association, 92, 1546-1554.

Lombardo, R., Durand, J.F. and De Veaux (2009), D., "Model building in multivariate additive PLS splines via the GCV criterion ". Journal of Chemometrics, vol. 23, issue 12, 607 - 617 .

Lombardo, R., Durand, J.F. and P. Leone, A. (2012), "Multivariate additive PLS spline boosting in Agro-Chemistry studies", Current Analytical Chemistry, vol. 8, issue 2, 236 - 253.

Wold, H. (1984), "PLS Regression", Encyclopaedia of Statistical Sciences. N. Johnson and S. Kotz eds. John Wiley, New York, Vol. 6, 58 -591.
