Partial Least-Squares Splines :

Presentation of the

PLSS package (Version 10.30)

J.F. Durand

Laboratoire de Probabilités et Statistique,
Montpellier II University, France

E-Mail: jf.durand@club-internet.fr
Web site: www.jf-durand-pls.com

I. Short presentation of the functions

1. Downloads

R-package ↦ http://cran.r-project.org/
Source files ↦ http://www.jf-durand-pls.com/ProgramSources.html

2. Four main R-functions

FUNCTION	TITLE	MENU
PLSL PLSS plscalibration	Linear PLS PLS through Splines Linear PLS for Near Infrared Spectroscopy	TITLE (0 to exit) 1: Bivariate Analysis for Spline Inputs (*1:) 2: Cross Validation 3: Generalized Cross Validation 4: A look at graphics for exploration and modeling (*4:) 5: Selection of the predictors (Remove/Add) (*5:) Selection:
MAPLSS	Multivariate Additive PLS Splines (**)	TITLE (0 to exit) 1: CV/GCV for pure main effects models (mandatory) 2: Automatic selection of interactions 3: Validation of the dimension and a look at the model 4: Pruning main effects and interactions Selection:

In "PLSL" : (*1:) "Bivariate Analysis"
(*4:) "Conversational PLS"
In "plscalibration" : (*1:) "Looking at the spectra"
(*4:) "Conversational PLS"
(*5:) not available
(**) MAPLSS involves bivariate interactions

3. Computational inputs

Common to all functions		
Input	*Description*	*Default*
X="matrix"	learning sample predictors (mandatory).	
Y="matrix"	learning sample responses (mandatory).	
Xtest="matrix"	test sample predictors.	missing (*)
Ytest="matrix"	test sample responses.	missing
D="vector"	stat. weights of observations, if D=1 equidistributed.	1
StandX="bool"	if =T, X is D-standardized, if =F, only D-centered.	T
StandY="bool"	idem for Y.	T
A="num"	number of PLS components.	1
prop="num"	proportion of left/predicted observations in CV.	0.1
GCV="num"	positive GCV parameter, if 0, CV is processed.	2

plscalibration		
Input	*Description*	*Default*
spect="vector"	integer vector of the training sample spectra.	1:nrow(X)
spectest="vector"	integer vector of the test sample spectra.	1:nrow(Xtest)
byscale="num"	integer value, 1/*byscale* means that only 1 column of X (wavelength) among *byscale* is accounted for.	1

MAPLSS		
Input	*Description*	*Default*
interaction="vector"	integer vector of predictors for possible interactions.	1:ncol(X)
PRESSprop="num"	threshold for accepting or not one interaction.	0.2

↦ All matrices must have non null dimnames.

↦ (*) Xtest=X in plscalibration.

4. Inputs for graphics

Common to all functions		
Input	*Description*	*Default*
bgpar="string"	background color name. Use colors( ) to choose.	"lightblue"
askpar="boolean"	if T, press ENTER to see next plot.	T
ptypar, cexpar, pchpar	usual R-plot parameters: pty, cex, and pch.	"s", 0.7, 1
typedata="boolean"	if T, residual plot with the name of the observations, if F, their number only.	T
qual="vector"	nrow(X) integers giving the color of the observations.	missing
names.qual="vector"	vector of *qual* level names to use in a legend sub-plot.	missing
titlepar="boolean"	if T, put a title in multiple plots.	T
colpar="num"	initialize the first color to be used.	1
prop="num"	proportion of left/predicted observations in CV.	0.1
matrow="num"	number of rows in a matrix of plots.	1
matcol="num"	number of columns in a matrix of plots.	1

"Looking at the spectra" in plscalibration		
Input	*Description*	*Default*
steps="boolean"	if T, displays the spectra one-by-one in *stepsorder* order.	F
stepsorder="boolean"	if T, the increasing order.	T
titlestring="string"	title of plots	"Training sample"
matrow, matcol	matrix dimension of the 3 plots: mean spectrum, first and second derivatives.	1, 3

3-D perspective plots in plscalibration and MAPLSS		
Input	*Description*	*Default*
thetapar, phipar, rpar	(θ, ϕ, r) eye spherical coordinates.	-60, 30,10

5. Spline inputs

Degree		
Input	*Description*	*Default*
degree="vector"	vector of ncol(X) integers, the degrees to transform the predictors, if degree=d, all predictors have the same degree, say d.	1
Knots		
Rapid strategy without a priori information		
Input	*Description*	*Default*
knots="vector"	vector of ncol(X) integers, the numbers of knots for the predictors, if knots=k, all predictors have the same number, say k.	0
equiknots="vector"	vector of ncol(X) T/F values. If T, equally spaced knots, if F, knots located at the *knots* - quantiles of the predictor. If equiknots=F, all predictors have knots at *knots* - quantiles.	F
Individual choice for each predictor		
Input	*Description*	*Default*
listknots="list"	list of ncol(X) vectors of the location of knots, if *listknots* is missing, then *knots* and *equiknots* are available.	missing

↦ Default values provide, for all predictors, degree=1 and knots=0 (empty set of knots) leading to a linear PLS model.

↦ The first item in the PLSS menu, "Bivariate Analysis for Spline Inputs", allows direct on-line graphical selection of the spline inputs.

II. Some real and simulated examples

Download the text file of these commands and the data sets involved, from the web-pages : Source Files and Data sets.
First input : the predictor matrix X *observations* × *variables*.
Second input : the response matrix Y *observations* × *variables*.

1. Linear PLS on Cornell Data

```
PLSL(cornell[,1:7],cornell[,8,drop=F],cexpar=0.8,matrow=3,matcol=3)
```

allows to preserve the matrix structure and the column name.

2. Linear PLS calibration on TECATOR Data

```
plscalibration(meatX,meatY[,1,drop=F],spect=1:172,Ytest= meatY[,1,drop=F],
spectest=173:215,matrow=3,matcol=1,titlestring= "Absorbance Calibration
Sample",askpar=T)
```

Training samples 1:172, test samples 173:215

```
plscalibration(meatX,meatY[,1,drop=F],spect=1:172,Ytest= meatY[,1,drop=F],
spectest=173:215,matrow=3,matcol=1,titlestring= "Calibration Sample",
steps=T,steporder=F,askpar=T)
```

Absorbance spectra are step by step sequentially displayed in descending order.

---

3. PLSS to model the satisfaction of health services

The survey made in the hospital of Aversa (Italy) was based on 235 patients asked on their satisfaction (5 degrees of valuation, from 1 to 5) concerning health care services.

```
aversaX=dget("aversaX.txt") # the 235x20 predictor matrix
attach(aversaX) # the ability of calling variables by their names
aversaY=dget("aversaY.txt") # the 235x1 response matrix
```

	Variables	Type	Meaning
personal info.	ETA	continuous	age
	TITOLODISTUDIO	categorical	education level
	PESO	continuous	weight in terms of cost of patients
	GD	continuous	days of recovery
	TARIFFA	continuous	cost of patients with respect to patology
tangibility	T1	categorical	instruments/machines
	T2	categorical	rooms
	T3	categorical	cleanliness
reliability	R1	categorical	time when patients call a sanitary operator
	R2	categorical	patients ask to be help by a sanitary operator
	R3	categorical	information received from sanitary operators
response capacity	CRis1	categorical	precise
	CRis2	categorical	fast
	CRis3	categorical	always available
assurance capacity	CRas1	categorical	let me feel trusting
	CRas2	categorical	necessary knowledge of sanitary operators
	CRas3	categorical	kind with me
	CRas4	categorical	good management
measure of empathy	E1	categorical	individual attention
	E2	categorical	understand my needs
response	Glob.Satisf.	categorical	global satisfaction

To obtain more easy interpretable results we regroup the 5 levels of the response into {1,2}, {3} and {4,5} that become the new response levels 1, 2 and 3.

```
aversaY3=dget("aversaY3.txt") # the 235x1 new response matrix.
```

Let `indicatorY3` *be the* 235×3 *binary indicator matrix of the levels that becomes the multi-response matrix. The function* `Bspline` *constructs the coding-matrix of variables through the use of B-spline basis functions, the input* `graph` *allows to see the B-splines :*

```
indicatorY3=Bspline(aversaY,degree=0,knots=2,equiknots=T,graph=T)
```

As a result of a preliminary campaign of PLSS tries based on splines of degree 0 and 2 equally spaced knots for all the predictors, eight predictors, namely 1, 2, 3, 4, 5, 11, 13 and 16, have a negligible influence on the responses and are therefore removed.

```
PLSS(aversaX[,c(6:10,12,14:15,17:20)],indicatorY3,degree=0,knots=2,
equiknots=T, prop=0.2,qual=Cras3,names.qual=paste("kindness",1:5))
```

```
 table of real and predicted training groups :
            real 1 real 2 real 3
predicted 1     26     17      3
predicted 2     12     60     18
predicted 3      0     23     76
```

With 20% of observations out at a time in the CV, the optimal PRESS is obtained by 2 components, PRESS(0.2,2)=2.1288. Colored 2-D maps of the individuals are obtained where each individual is colored according to his answer to the question Cras3 whose 5 levels measure the satisfaction of the service: kindness of the health operators. That service is revealed by the models as the most influential on the global satisfaction. The next table presents the classification of the predictors with respect to their respective rank of influence on the 3 responses, Glob.Satisf. 1, Glob.Satisf. 2 and Glob.Satisf. 3. The influence of a predictor on a response is measured by the range of the transformed values by the spline that is here a piecewise constant function.

	Classification of the services				
health service	*Glob.Satisf. 1*	*Glob.Satisf. 2*	*Glob.Satisf. 3*	Total	Class.
T1 (instr/machines)	9	1	1	11	2
T2 (rooms)	11	2	4	17	5
T3 (cleanliness)	12	10	9	31	11
R1 (time for help)	7	8	12	27	10
R2 (ask for help)	4	12	10	26	9
CRas1 (let me feel trusting)	3	4	5	12	3
CRas3 (kindness)	1	3	3	7	1
CRas4 (good management)	10	11	11	32	12
CRis1 (precise)	8	7	8	23	8
CRis3 (always available)	2	5	7	14	4
E1 (individual attention)	6	9	2	17	5
E2 (understand my needs)	5	6	6	17	5

As a result, a measure of the overall satisfaction of a service is proposed as the sum of its ranks of influence on the 3 responses.

---

4. Pure main effects model on Orange Juice Data

The building model stage

A campaign of PLSS tries allows to construct the main effects additive model.

```
try0<-PLSS(juicX,juicY[,1,drop=F])
```

Default spline parameters for all predictors: degree=1, knots=0.

```
try1<-PLSS(juicX,juicY[,1,drop=F],degree=2,knots=2)
```

Spline parameters for all predictors: degree = 2, 2 knots at quantiles.

```
try2<-PLSS(juicX,juicY[,1,drop=F],degree=2,knots=2,equiknots=T)
```

Spline parameters for all predictors: degree = 2, 2 equally spaced knots, by default knots at quantiles (equiknots=F).

```
localknots<-list(c(400,1600),c(10,20,40),c(10,40),c(2.5,5),c(160,400),
c(40,110), c(4,11,30),c(400,1700),c(100,300,500),c(600,2600))
try3<-PLSS(juicX,juicY[,1,drop=F],degree=2,listknots=localknots)
```

Spline parameters : degree = 2, individual knots for each predictor.

A look at the PLSS outputs

Results are stored in try3, a list of 5 objects:

try3$Xvariables, try3$degree, try3$knots, try3$equiknots and try3$listknots.

try3$Xvariables is the boolean indicator of the retained predictors in the selection of the main effects (Menu 5), the four others characterize the spline inputs.

```
try3
$Xvariables
[1] FALSE FALSE FALSE FALSE TRUE TRUE FALSE TRUE TRUE
[10] FALSE
```

```
$degree
[1] 2 2 2 2 2 2 2 2 2 2
$knots
[1] 0 0 0 0 0 0 0 0 0 0
$equiknots
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[10] FALSE
$listknots
$listknots[[1]] [1] 400 1600
$listknots[[2]] [1] 10 20 40
$listknots[[3]] [1] 10 40
$listknots[[4]] [1] 2.5 5.0
$listknots[[5]] [1] 160 400
$listknots[[6]] [1] 40 110
$listknots[[7]] [1] 4 11 30
$listknots[[8]] [1] 400 1700
$listknots[[9]] [1] 100 300 500
$listknots[[10]] [1] 600 2600
```

Recall that knots and equiknots are not available when listknots is non null.

Results of the final retained model

```
PLSS(juicX[,try3$Xvariables],juicY[,1,drop=F],degree=2,listknots=
localknots[try3$Xvariables])
```

Retained predictors :

```
dimnames(juicX)[[2]][try3$Xvariables] [1] "Ca" "Mg" "SO4" "HCO3"
```

degree : 2

knots location :

```
localknots[try3$Xvariables]
[[1]] [1] 160 400
[[2]] [1] 40 110
[[3]] [1] 400 1700]
[[4]] [1] 100 300 500
```

Optimal PRESS obtained with 2 components: "PRESS(0.1,2)=0.1157" with "prop=0.1", the proportion of observations out.

Optimal GCV : "GCV(1.8,2)=0.1144", with "GCV=1.8" the tuning GCV parameter.

---

5. MAPLSS for interactions on Orange Juice Data

Here are only searched eventual interactions between the main predictors 5, 6, 8 and 9, selected in the pure additive PLSS model.

```
MAPLSS(juicX,juicY[,1,drop=F],degree=2,listknots=localknots, GCV=1.8,
interaction=c(5,6,8,9))
```

Tuning parameter for the GCV criterion. When GCV=0, the Cross-Validation is processed, the default value is 2. One has to calibrate that parameter, that is, to find a dimension and a GCV-criterion value as close as possible to respectively the dimension and the PRESS-criterion value obtained in Cross-Validation.

Vector of predictors for the candidate interactions, the default, 1:ncol(X), means that all bivariate interactions are candidate: ncol(X)(ncol(X)-1)/2.*

No interaction is retained. Notice that testing all possible 45 interactions between the 10 predictors, gives the same result:

```
MAPLSS(juicX,juicY[,1,drop=F],degree=2,listknots=localknots, GCV=1.8)
```

---

6. Testing MAPLSS on simulated data

The function f constructs (X,Y) data bases. X, n by p, is a matrix of p n-dimensional covariate vectors uniformly generated in the unit hypercube. Y, n by 1, is the "signal + noise" response.

f to generate "Y=f(X)+noise" data		
Input	*Description*	*Default*
signal="string"	the signal, a function of the X columns.	""
n="integer"	number of observations.	100
p="integer"	number of predictors (X columns).	1
stdv="numerical"	standard deviation of the normal *N(0, stdv)* noise.	1
seedpar="numerical"	the seed for generating numbers at random.	20
Output	*Description*	*Default*
$X="matrix"	$n \times p$ training sample sample *U[0, 1]* predictors.	100 *U[0, 1]*
$Y="matrix"	$n \times 1$ training sample response.	100 *N(0, 1)*

```
f1<-"10*sin(pi*X[,1]*X[,2])+20*(X[,3]-0.5)^2+10*X[,4]+5*X[,5]"
```

Through the following R-commands, the response does not depend on the 5 last predictors (among 10) that contribute as pure noise:

```
experiment1<-f(f1,100,10)
X<-experiment1$X
Y<-experiment1$Y
```

```
MAPLSS(X,Y,degree=c(1,1,2,1,1,1,1,1,1,1),listknots=list(0.5,0.5,0.5,NULL,
NULL,NULL,NULL,NULL,NULL,NULL))
```

*The interaction detected is the true X1*X2, with a relative gain in the GCV of the pure additive main effects model that amounts to 53.96%.*

```
 2.2: Incorporating interactions step by step :
-------------------------------------------------
 Reference : Main effects GCV(2,2) = 0.1215937
-------------------------------------------------
candidate  1  :  X1*X2 ACCEPTED. at 20 % rel. GCV gain

      i j R2CRIT(2) PRESSCRIT   TOTCRIT A     GCV
X1*X2 1 2  0.076274 0.5291732 0.6054472 5 0.05598371
        %rel.GCVgain
X1*X2         53.96

candidate  2  :  X2*X4 NOT ACCEPTED. at 20 % rel. GCV gain

      i j R2CRIT(2)  PRESSCRIT    TOTCRIT A     GCV
X2*X4 2 4 0.0116159 0.05299399 0.06460989 5 0.05161322
        %rel.GCVgain
X2*X4          7.81
```