1 Introduction

L'objectif de ce rapport est de fournir un guide pour l'utilisateur de la régression Partial Least Squares linéaire (PLS) sous Splus. Il est l'aboutissement d'un Travail d'Etude et de Recherche (TER) ayant pour cadre l'option "Analyse Exploratoire de Données Multivariées" de la Maîtrise d'Ingénierie Mathématique de l'Université Montpellier II. Ce TER, proposé par J.F. Durand aux étudiants S. Roman et M. Vivien, s'est poursuivi par un stage dans le groupe de Biostatistique et Systèmes Dynamiques. Ce travail constitue la première partie d'un projet plus vaste présentant la modélisation PLS linéaire puis non linéaire additive (Durand et Sabatier 1997, Durand 1997).

Le langage Splus© utilisé pour les différentes fonctions **pls**, **plscv** ... est actuellement l'un des plus employés par les statisticiens développeurs. Ses fonctions statistiques à la pointe de l'actualité de la recherche le rendent de plus en plus attractif dans le monde de l'industrie. Le pré-requis Splus a été réduit le plus possible. L'utilisateur trouvera dans la Section 2 et dans l'Annexe 1 les premières notions du langage nécessaires à la saisie des données, les documents de référence étant le manuel d'utilisation MathSoft (1996) ainsi que l'introduction à Splus proposée par Baumgartner (1994).

La régression linéaire PLS (H. Wold 1966, S. Wold et al. 1983) très populaire en Chimiométrie, est de plus en plus appréciée par les autres secteurs industriels et économiques. La raison principale de son succès réside dans le fait qu'elle combine de façon efficace l'Analyse Exploratoire (réduction de dimension) et la modélisation (régression). L'enseignement de cette méthode fait maintenant partie des programmes du second cycle universitaire à l'Université Montpellier II. Pour des références en Français, citons Tenenhaus et al. (1995), Bry (1996) et Tenenhaus (1998).

La Section 2 est une introduction au logiciel Splus, utilisé sous UNIX, à propos de l'exploration du jeu de données sur l'OCDE (Bertier et Bouroche 1975). Cet exemple servira dans la Section 3 à montrer les différentes possibilités des programmes PLS écrits sous Splus par J.F. Durand et R. Sabatier. Enfin, l'Annexe est constituée de trois parties. La première est un résumé des principales fonctions Splus permettant la saisie des données. La présentation des fonctions Splus non natives à ce langage, utilisées dans ce travail constitue l'Annexe 2. La dernière partie de l'Annexe est une synthèse des principales propriétés mathématiques de la méthode.

2 Exploration des données sous Splus

La première manipulation pour pouvoir utiliser le logiciel de statistique Splus sous UNIX, consiste, lorsque l'on est dans un répertoire, à créer un sous-répertoire que l'on nommera **.Data**. C'est dans ce sous-répertoire que se trouvent tous les objets (matrices, fonctions,...) créés sous Splus. Pour lancer le logiciel Splus, il faut taper :

```
Splus
```

Pour quitter Splus,

```
> q()
```

Le "prompt" > indique que l'on travaille sous Splus. Si on utilise à nouveau Splus dans le même répertoire, on n'a pas besoin de recréer le sous-répertoire **.Data**. Pour visionner la liste des objets créés, il faut faire appel à la fonction **ls** :

```
> ls()
```

Le logiciel Splus est composé d'une fenêtre de dialogue et d'une fenêtre graphique pour laquelle il peut s'avérer utile de définir des couleurs. Pour cela, mais aussi pour que la fenêtre graphique, X11() ou motif(), s'ouvre automatiquement lors du lancement de Splus, il est pratique d'écrire dans un éditeur de texte la fonction **.First**, qui est reconnue et exécutée automatiquement par Splus lors de l'appel du logiciel :

```
.First<-function()
{X11("-xrm ' Splus*colors: white black 2 red 2 green 2 blue 2 orange' ")}
```

ou bien :

```
.First<-function()
{motif("-xrm ' sgraphMotif*colors: black orange magenta red yellow2 turquoise1 green2
pink2 green4 blue ' ")}
```

Après avoir sauvé ce programme sous le nom .First, il faut le compiler sous Splus à l'aide de la fonction **source** :

```
>source(".First")
```

2.1 Premiers traitements

On suppose que l'utilisateur dispose des données sous forme de 2 matrices ocdeX et ocdeY. Dans le cas contraire, voir l'Annexe 1 qui décrit comment rentrer des données sous Splus.

```
Résumé des principales commandes :
>xgobi(ocde)
>as.data.frame(ocde)
>attach(ocde)
>Dvar(ocde,cor=T)
>plot(IMP,EXP)
>abline(lm(EXP ~ IMP))
>detach()
```

Les données

Les tableaux de données ocdeX et ocdeY contiennent respectivement les mesures de 13 variables socio-économiques et 5 variables de consommation recueillies en 1970 sur les 18 pays de l'OCDE. Ces variables sont des variables quantitatives, voir Bertier et Bouroche (1975).

```
>ocdeX
```

	POP	DENS	POPG	AGRF	INDU	GNP	GDPA	FCF	RR	OFR	DR	IMP	EXP
D	60848	245	1.05	9.6	49.1	2520	3.6	24.4	37.9	10940	6.50	24926	29052
A	7373	88	0.50	19.1	39.9	1690	7.0	23.2	37.5	1563	5.00	2825	2412
B	9984	332	0.60	5.4	44.8	2352	5.4	23.1	35.1	2406	7.00	9984	10069
CDN	21089	2	1.85	8.2	32.3	3460	5.9	21.7	35.2	3846	6.00	13137	13754
DK	4893	114	0.75	11.9	38.5	2860	8.9	22.0	37.1	384	9.00	3800	2958
E	32949	65	0.95	30.7	37.1	870	15.0	22.0	22.4	1518	6.50	4233	1900
USA	203213	22	1.35	4.6	33.7	4660	2.9	16.7	31.5	12306	5.75	36052	37988
FI	4706	14	0.70	24.5	34.6	1940	14.7	23.8	35.9	379	6.00	2023	1985
F	50325	91	1.05	15.1	40.6	2770	6.0	25.4	38.1	4617	7.50	17392	15020
G	8866	67	0.70	48.2	22.5	950	20.3	29.7	26.9	290	6.50	1594	554
IRL	2921	42	0.25	28.4	29.7	1040	19.7	19.9	30.7	694	7.31	1413	891
I	54120	180	0.85	21.5	43.1	1520	11.3	20.5	33.3	4642	5.50	12450	11729
JAP	102380	277	1.05	18.8	35.0	1630	8.7	35.2	21.2	3072	6.00	15024	15990
N	3851	12	0.80	14.7	36.8	2530	6.5	25.3	43.4	607	4.50	2943	2203
NL	12873	352	1.25	7.5	41.3	2190	7.0	25.5	41.9	2621	6.00	10991	9965
P	9583	105	0.90	31.5	35.5	600	17.7	18.4	24.0	1442	3.50	1231	823
UK	55643	228	0.65	2.9	46.8	1970	3.0	17.3	39.0	2469	7.00	19956	17515
S	7969	18	0.70	8.8	40.4	3230	5.9	23.6	48.1	506	7.00	5899	5698

```
>ocdeY
```

	CAL	LODG	ELEC	EDUC	TV
D	2990	8.6	3322	3.0	231
A	2990	6.6	2647	4.4	134
B	3150	5.0	2814	5.3	184
CND	3160	8.2	8199	5.7	279
DK	3180	9.0	2413	6.0	244
E	2750	6.4	1245	2.1	84
USA	3210	7.7	7013	5.1	392
FI	2900	7.9	3836	6.3	193
F	3160	8.2	2407	4.8	185
G	2910	10.1	823	2.4	9
IRL	3450	4.0	1577	4.2	111
I	2940	5.1	1810	5.8	146
JAP	2460	11.9	2734	4.5	190
N	2910	8.8	12976	5.8	175
NL	3240	9.7	2565	6.7	197
P	2930	4.3	607	1.4	29
UK	3180	7.7	3680	4.2	263
S	2750	13.4	6803	7.4	288

Abréviation des variables et des pays

Variables		*Pays*	
POP	population (en 1000 hab.)	D	Allemagne
DENS	densité au km^2	A	Autriche
POPG	taux d'accroissement de la pop.	B	Belgique-Luxembourg
AGRF	% d'actifs dans l'agricul. et la pêche	CND	Canada
INDU	% d'actifs dans l'industrie	DK	Danemark
GNP	PNB en $ par habitants	E	Espagne
GDPA	% du PIB en agriculture	USA	États-Unis
FCF	formation du capital fixe en % du PNB	FI	Finlande
RR	recettes courantes en % du PNB	F	France
OFR	réserves officielles (en millions $)	G	Grèce
DR	taux d'escompte	IRL	Irlande
IMP	importations (en millions $)	I	Italie
EXP	exportations (en millions $)	JAP	Japon
CAL	calories par habitant et par jour	N	Norvège
LODG	nombre de logements pour 1000 hab.	NL	Pays-Bas
ELEC	consommation d'électricité en kWh/habi./an	P	Portugal
EDUC	dépenses publiques dans l'éducation en % du PNB	UK	Grande-Bretagne
TV	nombre de TV pour 1000 habi.	S	Suède

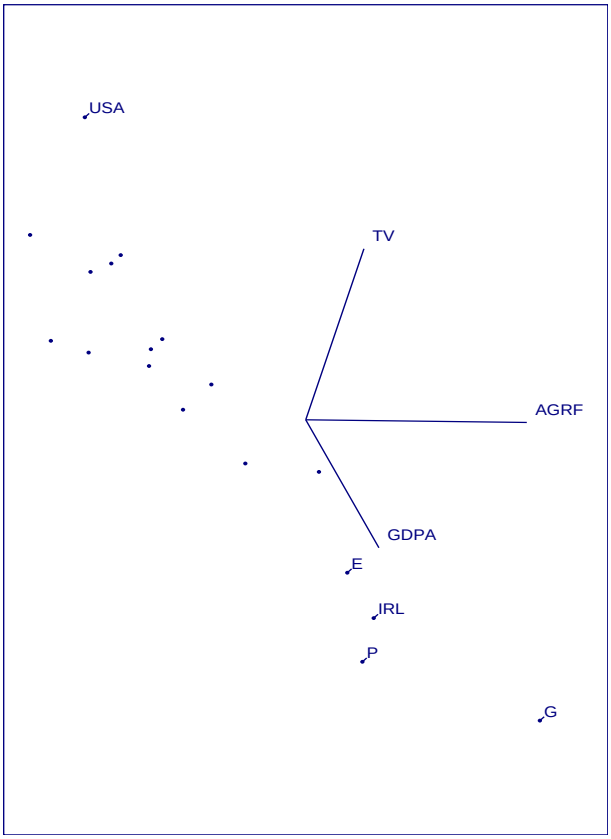
L'objectif est d'explorer ces données, notamment d'étudier les liaisons linéaires entre les variables et d'obtenir des cartographies des individus permettant de repérer d'éventuels pays se distinguant particulièrement des autres et de former des groupes de pays ayant des caractéristiques similaires.

Avant l'analyse, il convient de faire une première exploration graphique des données et d'étudier la matrice des corrélations. Cela permet de détecter d'éventuelles erreurs de saisies, de prendre connaissance des données, des ordres de grandeur, de commencer une première interprétation.

Sous Splus, la construction de ces graphiques peut se faire avec le programme **xgobi**© (Swayne, Cook, Buja 1991) qui est un logiciel d'exploration graphique de données multivariées aux multiples possibilités. Seule l'option de représentation tridimensionnelle est utilisée ici. Si **xgobi** n'est pas disponible, la fonction **brush** de Splus permet aussi ce type de représentation. L'argument d'entrée de la fonction **xgobi** étant une matrice, on fabrique d'abord la matrice ocde à partir de ocdeX et ocdeY.

```
>ocde<-cbind(ocdeX,ocdeY)
>xgobi(ocde)
```

Le lancement de cette fonction fait apparaître à l'écran une fenêtre interactive, dans laquelle se trouvent des menus permettant de sélectionner le type de graphiques, pour les variables, que l'on désire : en une, deux, trois dimensions ou même plus.
D'une manière générale, le bouton gauche et le bouton du milieu de la souris permettent de sélectionner des objets, bouger les graphiques etc, et le bouton droit permet d'avoir une fenêtre d'aide.
Le menu Identify permet d'identifier les labels des individus sur tous les plots. Il suffit alors de promener la souris sur le graphique et de cliquer sur les individus qui nous intéressent.
Le menu Rotate permet de représenter les individus dans un repère à trois axes représentant trois variables.



On remarque que les USA sont opposés au groupe Grèce, Irlande, Espagne et Portugal

dans le repère GDPA, AGRF, TV.

Pour quantifier ces corrélations, c'est-à-dire construire la matrice des corrélations, on utilise la fonction **Dvar**, fonction programmée par R.Sabatier et J.F.Durand. Elle permet de pondérer les individus par des poids qui par défaut sont équidistribués.

```
>round(Dvar(ocde,cor=T)$V,2)
```

	POP	DENS	POPG	AGRF	INDU	GNP	GDPA	FCF	RR
POP	1.00	0.06	0.42	-0.32	0.05	0.49	-0.41	-0.11	-0.26
DENS	0.06	1.00	-0.01	-0.32	0.57	-0.17	-0.30	0.26	-0.03
POPG	0.42	-0.01	1.00	-0.33	-0.03	0.52	-0.39	0.04	-0.05
AGRF	-0.32	-0.32	-0.33	1.00	-0.68	-0.75	0.92	0.27	-0.61
INDU	0.05	0.57	-0.03	-0.68	1.00	0.19	-0.67	-0.22	0.47
GNP	0.49	-0.17	0.52	-0.75	0.19	1.00	-0.76	-0.19	0.53
GDPA	-0.41	-0.30	-0.39	0.92	-0.67	-0.76	1.00	0.09	-0.58
FCF	-0.11	0.26	0.04	0.27	-0.22	-0.19	0.09	1.00	-0.16
RR	-0.26	-0.03	-0.05	-0.61	0.47	0.53	-0.58	-0.16	1.00
OFR	0.81	0.16	0.51	-0.43	0.31	0.55	-0.52	-0.22	-0.02
DR	-0.06	0.11	-0.17	-0.22	0.11	0.24	-0.13	0.05	0.20
IMP	0.87	0.26	0.52	-0.60	0.37	0.64	-0.68	-0.19	0.06
EXP	0.86	0.25	0.53	-0.59	0.35	0.65	-0.67	-0.16	0.06
CAL	-0.05	0.00	-0.03	-0.30	0.05	0.26	-0.10	-0.59	0.31
LODG	0.10	0.00	0.25	-0.19	-0.06	0.38	-0.33	0.61	0.33
ELEC	0.12	-0.39	0.33	-0.47	0.00	0.64	-0.51	-0.05	0.53
EDUC	-0.05	0.00	0.11	-0.59	0.21	0.60	-0.47	0.04	0.71
TV	0.55	-0.01	0.43	-0.86	0.35	0.92	-0.81	-0.25	0.50

	OFR	DR	IMP	EXP	CAL	LODG	ELEC	EDUC	TV
POP	0.81	-0.06	0.87	0.86	-0.05	0.10	0.12	-0.05	0.55
DENS	0.16	0.11	0.26	0.25	0.00	0.00	-0.39	0.00	-0.01
POPG	0.51	-0.17	0.52	0.53	-0.03	0.25	0.33	0.11	0.43
AGRF	-0.43	-0.22	-0.60	-0.59	-0.30	-0.19	-0.47	-0.59	-0.86
INDU	0.31	0.11	0.37	0.35	0.05	-0.06	0.00	0.21	0.35
GNP	0.55	0.24	0.64	0.65	0.26	0.38	0.64	0.60	0.92
GDPA	-0.52	-0.13	-0.68	-0.67	-0.10	-0.33	-0.51	-0.47	-0.81
FCF	-0.22	0.05	-0.19	-0.16	-0.59	0.61	-0.05	0.04	-0.25
RR	-0.02	0.20	0.06	0.06	0.31	0.33	0.53	0.71	0.50
OFR	1.00	-0.06	0.91	0.94	0.16	-0.03	0.13	-0.08	0.54
DR	-0.06	1.00	0.09	0.06	0.32	0.23	-0.22	0.26	0.29
IMP	0.91	0.09	1.00	0.99	0.17	0.12	0.18	0.07	0.70
EXP	0.94	0.06	0.99	1.00	0.15	0.13	0.20	0.06	0.70
CAL	0.16	0.32	0.17	0.15	1.00	-0.47	-0.02	0.15	0.22
LODG	-0.03	0.23	0.12	0.13	-0.47	1.00	0.33	0.39	0.37
ELEC	0.13	-0.22	0.18	0.20	-0.02	0.33	1.00	0.50	0.55
EDUC	-0.08	0.26	0.07	0.06	0.15	0.39	0.50	1.00	0.62
TV	0.54	0.29	0.70	0.70	0.22	0.37	0.55	0.62	1.00

En fait, cette fonction calcule la matrice des covariances ou des corrélations de la matrice donnée, suivant que **cor** égale F ou T. De plus, cette fonction retourne une liste composée de la matrice des corrélations, de la matrice centrée-réduite, du vecteur des moyennes, du vecteur des variances. **$V** signifie que l'on veut l'élément V de la liste, la fonction **round** arrondit, ici à 2 chiffres après la virgule, les éléments de V.

Cette matrice permet de repérer les variables qui sont linéairement corrélées, par exemple IMP et EXP (0.99).

La fonction **plot** va être utilisée pour visualiser la relation linéaire entre ces variables. Tout d'abord, il est utile de transformer la matrice ocde en data.frame pour pouvoir utiliser le nom des variables (les colonnes de ocde).

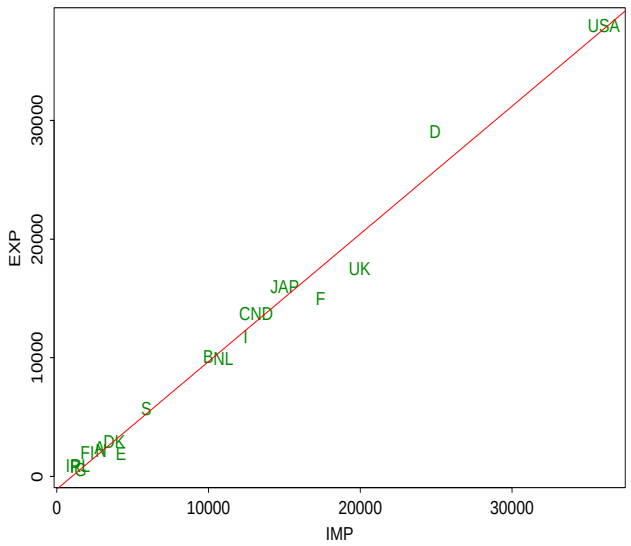
```
>ocde<-as.data.frame(ocde)
> attach(ocde)
```

On peut maintenant utiliser la fonction **plot** des deux manières suivantes :

```
>plot(IMP,EXP)
>plot(ocde[,12],ocde[,13])
>abline(lm(EXP ~ IMP))
```

Cette écriture ne permet pas d'afficher le nom des individus sur le graphique. Pour ce faire, si l'on désire de plus la couleur :

```
>plot(IMP,EXP,type=''n'')
> text(IMP,EXP,dimnames(ocde)[[1]],col=9)
>abline(lm(EXP ~ IMP),col=4)
>detach()
```



En vue d'effectuer la régression linéaire multiple de variables dites réponses sur les variables dites explicatives, ici, variables de consommation sur les variables socio-économiques, la fonction **lm** peut être mise en oeuvre.
On propose de faire la régression de la variable CAL, centrée et réduite, sur les variables de ocdeX centrées et réduites.
Commençons donc par centrer et réduire les données. Nous avons vu précédemment que la fonction **Dvar** permettait de l'obtenir.

```
>ocdeXcr<-Dvar(ocdeX,cor=T)$U
>ocdeYcr<-Dvar(ocdeY,cor=T)$U
```

Maintenant, il est possible d'utiliser la fonction **lm**. La variable CAL étant la première variable (colonne) de ocdeYcr, il faut taper :

```
>lm(ocdeYcr[,1] ~ ocdeXcr)
```

Si l'on préfère manipuler des data.frame, il convient d'utiliser les fonctions **as.data.frame**, **attach**, **lm** de la manière suivante :

```
>ocdeXcr<-as.data.frame(ocdeXcr)
>ocdeYcr<-as.data.frame(ocdeYcr)
>attach(ocdeXcr)
>attach(ocdeYcr)
>lm(CAL ~ POP+DENS+POPG+AGRF+INDU+GNP+GDPA
     +FCF+RR+OFR+DR+IMP+EXP)
>detach()
>detach()
```

```
Call :
lm(formula = CAL ~ POP + DENS + POPG + AGRF + INDU + GNP + GDPA + FCF + RR +
OFR + DR + IMP + EXP)
```

```
Coefficients :
  (Intercept)       POP      DENS       POPG      AGRF      INDU        GNP
8.365414e-16 -1.238146 0.3207027 -0.5018269 -1.555928 -1.466598 -0.3299705
      GDPA        FCF         RR       OFR        DR       IMP       EXP
0.2815631 -0.2997332 0.03477639 2.209719 0.1105386 4.557707 -5.258492

Degrees of freedom : 18 total; 4 residual
Residual standard error : 0.4179749
```

(Intercept) est le coefficient constant du modèle, ici, on peut le considérer comme nul car les variables sont centrées et réduites.

On peut voir que les variables IMP et EXP ont des coefficients de signes opposés. Cela voudrait dire que ces deux variables n'influent pas sur CAL dans le même sens. Or ces deux variables sont corrélées à 0.99. Il semblerait donc logique qu'elles aient un coefficient de même signe. Ce phénomène bien connu en régression linéaire plaide pour l'utilisation de méthodes faisant appel à la réduction de dimension comme la régression sur composantes principales, ou mieux, la régression Partial Least Squares.

2.2 Analyse en Composantes Principales

On rappelle brièvement que l'analyse en composantes principales est une méthode d'analyse descriptive : l'objectif est de représenter sous forme graphique le maximum de l'information contenue dans un tableau de données quantitatives.

Le but est de réduire l'espace de représentation des données pour une meilleure compréhension. Pour réduire l'espace, on construit des variables synthétiques, appelées composantes principales, résumant l'information : ce sont des combinaisons linéaires des variables initiales. Les coordonnées d'une composante principale indiquent la position des projections du nuage des individus sur l'axe principal associé.

Les composantes principales sont ordonnées par valeurs décroissantes de leurs variances. Ces dernières fournissent la dispersion des individus projetés sur les axes principaux. Elles sont obtenues par les valeurs propres de la matrice des variances-covariances ou des corrélations si les variables sont réduites.

Les premières composantes sont retenues, ce sont celles qui expliquent la plus grande

part de l'information, les autres sont négligées, elles sont considérées comme du "bruit". La fonction **prcomp** native sous Splus ne fournit qu'une version élémentaire de la méthode.

La fonction **acpxqd**, programmée par R.Sabatier, permet de faire une ACP généralisée associée à un triplet (X,Q,D) où Q est une métrique sur l'espace des lignes et D la matrice diagonale des poids des individus. Pour une ACP usuelle, Q=I_p et D=$n^{-1}I_n$, si n est le nombre d'individus ou lignes de X. Cette option est l'option par défaut, d'autres choix permettent de retrouver les principales méthodes de l'Analyse de Données (Analyse des Correspondances simple et multiple, Analyse Discriminante, etc).

```
>acpxqd(ocde)
```

Ce programme dispose d'une interface conversationnelle.

```
                    - D.S.D. d'un triplet (X,Q,D) -

                          donnees centrees et reduites

                           moyenne des variables de X

        POP       DENS       POPG       AGRF      INDU       GNP       GDPA
moy  36310.33  125.2222  0.8861111    17.3    37.87222  2154.556  9.416667
        FCF        RR         OFR        DR        IMP       EXP        CAL
moy  23.20556    34.4     3016.778  6.253333  10326.28  10028.11  3014.444
       LODG      ELEC       EDUC        TV
moy  7.922222  3748.389  4.727778  185.2222
                          variance des variables de X

         POP        DENS       POPG       AGRF      INDU       GNP        GDPA
var  2358833222  12482.84  0.1213349  133.0889   38.2809    1003491    30.35694
         FCF        RR         OFR        DR        IMP        EXP        CAL
var   18.12052  49.90444   11192241  1.409633  87229411  102981809  49469.14
        LODG       ELEC       EDUC        TV
var  5.858395  9244601  2.552006  8301.062
```

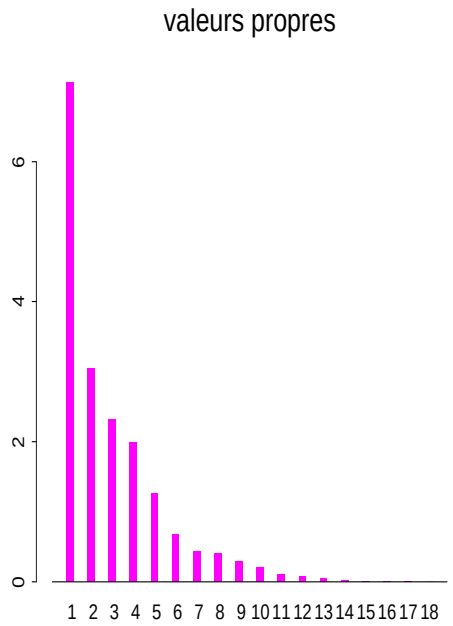
Ce tableau donne les moyennes et les variances de chacune des variables.

```
Inertie totale de (X,Q,D) = 18
```

L'inertie totale est de 18 (ce qui est égal aux nombre de variables) car l'ACP est centrée-réduite.

```
barplot des valeurs propres (o ou n) ?
```

1 : o



Le diagramme en batons des valeurs propres peut permettre de choisir le nombre d'axes résumant correctement l'information contenue dans le nuage de points. On choisit la valeur où la courbe dessinée par le diagramme présente une rupture de pente. Ici, on prendrait 5 ou 6 axes. Mais il existe d'autres critères de choix du nombre d'axes. Le tableau suivant donne les valeurs propres, le pourcentage d'inertie qu'elles représentent, le pourcentage d'inertie cumulé et le coefficient RV pour la reconstruction du tableau XX'.

```
    val.pro.  % inert.  % cumul.      RV
1    7.1390     39.66     39.66   0.8410195
2    3.0413     16.90     56.56   0.9141572
3    2.3151     12.86     69.42   0.9539747
4    1.9818     11.01     80.43   0.9821272
5    1.2655      7.03     87.46   0.9933774
6    0.6738      3.74     91.20   0.9965438
7    0.4339      2.41     93.61   0.9978542
8    0.4126      2.29     95.90   0.9990373
9    0.2847      1.58     97.48   0.9996003
10   0.1995      1.11     98.59   0.9998766
11   0.1016      0.56     99.15   0.9999483
12   0.0695      0.39     99.54   0.9999818
13   0.0451      0.25     99.79   0.9999960
14   0.0213      0.12     99.91   0.9999991
15   0.0107      0.06     99.97   0.9999999
16   0.0037      0.02     99.99   1.0000000
17   0.0006      0.00     99.99   1.0000000
18   0.0000      0.00     99.99   1.0000000
```

Un deuxième critère consiste à ne garder que les valeurs propres supérieures à 1, qui est l'inertie moyenne que peut porter un axe. Avec ce critère on choisirait 5 axes.
Le dernier critère porte sur le cumul de l'inertie des valeurs propres. On se fixe un seuil et on choisit le nombre d'axes nécessaires pour atteindre ce seuil. La plupart du temps ce seuil est fixé aux alentours de 85 %.
On a donc choisi 5 axes mais nous n'en interpréterons que 3.
Remarque : La fonction **acpxqd** ne retient par défaut que trois axes. Pour choisir un nombre d'axes supérieur, il faut préciser le paramètre **k** ($k \leq p$ où p est le nombre de variables).

```
>ocdeacp<-acpxqd(ocde,k=6)
```

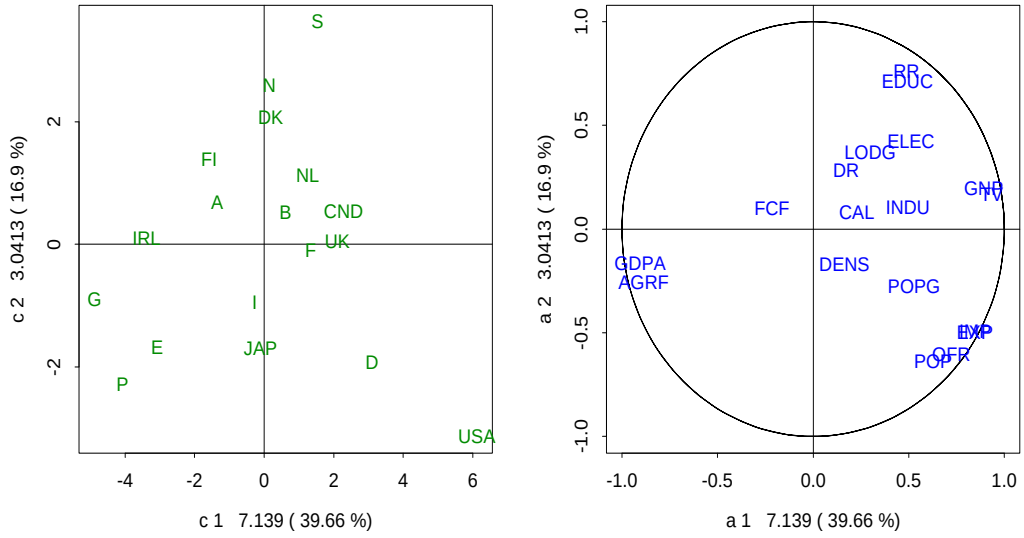
Les sorties sont stockées dans `ocdeacp`.

```
aides a l'interprétation pour les u.s.  (o/n) ?
1 :  n
```

```
aides à l'interpretation pour les variables (o/n) ?
1 : n
```

Ces aides à l'interprétation permettent d'afficher les contributions absolues et relatives des individus et des variables aux axes principaux.

```
 graphique pour les u.s.  et les variables (o/n) ?
1 : o
axe horizontal (<=6) ?
1 : 1
axe vertical (<= 6 ) ?
1 : 2
```



Les résultats graphiques ont permis à Bertier et Bouroche (1975) d'interpréter les trois premières composantes principales de la façon suivante :

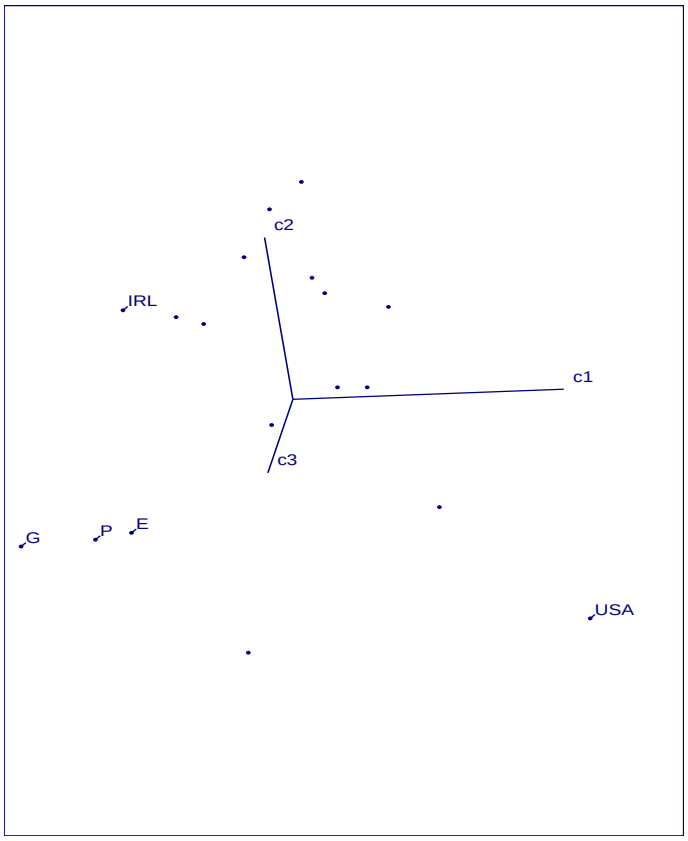
L'axe 1 opposant le PNB par tête au pourcentage du PIB dans l'agriculture peut s'interpréter comme le facteur *revenu par tête*. Il n'explique que 39.7% de l'inertie totale.

Le deuxième axe peut être considéré comme *l'importance des dépenses publiques*. Il explique encore 16.40% de l'inertie totale.

Le troisième axe qui n'explique plus que 12.9% de l'inertie totale a deux dénominations qui conviennent : *jeunesse* ou *investissement*.

On peut représenter le nuage des points dans le repère des trois premières composantes

principales (par exemple), en utilisant la fonction **xgobi** sur la matrice de ces composantes obtenue grâce à `ocdeacp$C`. Cela permet une vision plus exacte de la proximité entre pays que par l'examen des plans principaux.



3 La régression PLS linéaire

```
Résumé des principales commandes
>pls(ocdeX,ocdeY)
>ocdeplscv<-plscv(ocdeX,ocdeY,A=13,prop=0.05)
>plscv.plot(ocdeplscv)
>pls(ocdeXa,ocdeYa,Xtest=ocdeXt,Ytest=ocdeYt)
```

3.1 Motivations pour la régression PLS

Nous avons vu sur l'exemple les limites de la régression linéaire multiple (OLS) lorsque certains prédicteurs sont fortement corrélés.
Ainsi la modélisation de la variable CAL obtenue par OLS

$\widehat{CAL} = -1.238POP + 0.320DENS - 0.501POPG - 1.555AGRF - 1.466INDU - 0.329GNP + 0.281GDPA - 0.299FCF + 0.034RR + 2.209OFR + 0.110DR + 4.557IMP - 5.258EXP$

semble de qualité douteuse si l'on examine les coefficients les plus élevés associés aux variables IMP et EXP (4.557 et -5.258 respectivement). Ces variables étant fortement corrélées positivement (0.99), il semble douteux qu'elles puissent contribuer de façon opposée dans la modélisation de la variable CAL.

En outre, ce jeu de données se prête mal à l'utilisation de la régression aux moindres carrés ordinaires pour une autre raison : peu d'observations (18) ont été mesurées par rapport au nombre de variables explicatives (13). Il est donc nécessaire, si l'on désire modèliser linéairement les variables de consommation, d'utiliser une méthode mettant en oeuvre une étape de réduction de dimension à partir de composantes principales non corrélées.

La Régression sur Composantes Principales (RCP) est une des plus utilisées dans la pratique. Elle présente cependant le désavantage de régresser les réponses $Y_{n\times q}$ sur des résumés des prédicteurs $X_{n\times p}$ fournis par les composantes principales de plus grande variance qui ne sont pas forcément les plus explicatives des réponses.

La régression Partial Least Squares, PLS, (Wold 1966) quant à elle, construit des composantes principales qui résument X "dans le même temps" qu'elles expliquent les réponses. Aussi, PLS qui procède par projection des réponses sur les composantes principales comme le fait RCP, est maintenant préférée à cette dernière.

La fonction **pls** programmée sous Splus par Durand et Sabatier, permet une modélisation linéaire multiréponses. Cette possibilité peut-être intéressante lorsque les réponses sont mutuellement corrélées. Il est clair que cette fonction peut aussi modèliser séparément chaque réponse. Les modèles obtenus pour une réponse donnée par l'une ou l'autre approche sont généralement différents sauf si le nombre de composantes principales PLS est pris égal au rang de X. Dans ce cas en effet, le modèle PLS est identique au modèle OLS (on sait que le modèle OLS, $\widehat{Y} = P_X Y$ où P_X est le projecteur usuel, fournit le même résultat qu'il soit utilisé en uniréponse ou en multiréponses).

Une des particularités de PLS est que le modèle obtenu dépend du nombre A de composantes prises en compte. Ce nombre de composantes est donc un super-paramètre qu'il faudra choisir avec soin. Comme nous l'avons déjà indiqué, PLS(X,Y)≡OLS(X,Y)

lorsque $A = rang(X)$, ce qui constitue une des propriétés attractives de cette méthode.

Enfin, dernière propriété intéressante de PLS,

$$PLS(X, X) \equiv ACP(X).$$

C'est-à-dire, si la matrice des réponses Y est égale à la matrice X des prédicteurs, les composantes principales PLS sont identiques aux composantes principales usuelles.

Ces propriétés qui font de PLS à la fois un outil de modélisation et d'exploration des données expliquent sa popularité grandissante auprès d'un public d'utilisateurs à la recherche de méthodes robustes et efficaces.

3.2 La régression PLS sous Splus

La fonction **pls** dispose d'une interface conversationnelle. Dans sa version la plus simple (sans option), elle centre et réduit les matrices des variables explicatives et réponses, ocdeX et ocdeY :

```
>pls(ocdeX,ocdeY)
```

```
                              PLS lineaire
 Variance totale de Y=5
Variance totale de X=13
```

Puisque PLS centre et réduit sauf indication contraire, la variance totale des prédicteurs est égale à 13 et celle des réponses à 5.

```
Dimension 1
   cov(t,u)= 2.73487 r(t,u)= 0.826 stdev(t)= 2.237302 stdev(u)= 1.48067

                CAL  LODG   ELEC   EDUC    TV  % VAR Y
    R2 part.  0.071  0.08  0.253  0.241  0.85   29.885

% VAR X expliquee par la comp.  = 42.894224317400
```

```
 nombres d'axes supplementaires ?
1 :  1
```

La réponse est cruciale pour PLS, puisque le nombre total A d'axes ou de composantes (la dimension du modèle) est le super-paramètre de la méthode. Ce nombre A ne doit pas excéder le nombre de variables explicatives $p = 13$ (dans ce cas PLS = OLS et les composantes forment une base de l'espace des prédicteurs). Lors de la première exécution de la méthode, il est utile de choisir chaque fois un axe supplémentaire pour voir évoluer les deux critères de reconstruction de la variance des prédicteurs et des réponses (pour l'axe 1, % VAR X = 42.89 et % VAR Y = 29.88). La stratégie consiste à s'arrêter lorsque % VAR X est suffisamment grand, pour un gain faible dans % VAR Y. Ainsi ce critère basé sur l'ajustement des données conduit à $A = 6$. On verra dans la Section suivante d'autres critères de validation du nombre d'axes basés sur la prédiction.

```
 Dimension 6
cov(t,u)= 0.2705583 r(t,u)= 0.594 stdev(t)= 0.59378 stdev(u)= 0.7669466

                CAL    LODG    ELEC    EDUC      TV  % VAR Y
    R2 part.  0.169   0.021   0.005   0.003   0.010    4.152
     R2 cum.  0.729   0.710   0.769   0.651   0.905   75.280

% VAR X expliquee par la comp.  = 3.42084558543426
% VAR X expliquee = 93.7690576953179
```

Si l'on choisit le nombre maximum d'axes, $A = 13$, la régression PLS s'identifie à la régression aux moindres carrés usuelle pour les cinq variables réponses.

```
 Dimension 13
cov(t,u)= 0.02239014 r(t,u)= 0.84 stdev(t)= 0.05578502 stdev(u)=
0.477793
```

```
              CAL    LODG    ELEC    EDUC      TV   % VAR Y
 R2 part.   0.047   0.095   0.000   0.000   0.018     3.222
  R2 cum.   0.961   0.927   0.814   0.877   0.993    91.449

% VAR X expliquee par la comp.  = 0.0239382230611417
% VAR X expliquee = 100
```

A chaque étape, sont affichés

- **cov(t,u)** est la covariance entre la composante principale t associée à X, et u associée à Y,
- **r(t,u)** est le coefficient de corrélation simple entre t et u,
- **stdev(t)** et **stdev(u)** sont les écarts-types de t et u.

La covariance entre t et u, $cov(t,u) = r(t,u)stdev(t)stdev(u)$, est le critère à maximiser pour obtenir les composantes PLS optimales. C'est un compromis entre le critère de l'Analyse des Corrélations Canoniques et celui de l'ACP sur les tableaux des variables explicatives et des réponses.
D'autre part, pour chaque réponse, on obtient la valeur du coefficient R^2 de la régression partielle entre la réponse et la composante t. De même, R^2cum correspond au coefficient de corrélation multiple entre la réponse et l'espace engendré par les composantes.
On voit ici que la variable TV est très bien expliquée par t_1 ($R^2part = 0,85$). On pourra le vérifier graphiquement plus loin.

```
nombre d'axes supplementaires ?
1 :  0
Modeles PLS suivant le Nb de Composantes
1 :  Sur variables centrees reduites
2 :  Sur variables initiales
Selection :  1
```

Répondre 0 à `Selection :` permet de ne pas afficher les modèles.

```
[[6]] :
              POP         DENS         POPG         AGRF         INDU          GNP
 CAL   -0.43498319   0.45515689   0.17905210  -0.15214652  -0.49262315   0.09986199
LODG    0.17625858  -0.18600008   0.08873576  -0.03694422  -0.01680691   0.22956052
ELEC    0.07087953  -0.43201489   0.10335657  -0.16774508  -0.01733882   0.29204712
EDUC   -0.11764383  -0.05063768   0.09090961  -0.20230780  -0.09613038   0.27698157
  TV    0.12271461  -0.17497891   0.05313857  -0.17417230   0.01339607   0.27193105
             GDPA          FCF          RR          OFR          DR          IMP
 CAL    0.2830619  -0.76622161   0.2368949   0.24886102   0.3292480   0.08549177
LODG   -0.2081633   0.71014832   0.1438327  -0.21320685   0.1220921  -0.03467954
ELEC   -0.2611244   0.19022071   0.2527434  -0.13262225  -0.3396924  -0.07315540
EDUC   -0.1504819   0.14674231   0.3626527  -0.12680395   0.1120486  -0.05614414
  TV   -0.1631052  -0.01596207   0.1686109   0.01354928   0.1659059   0.09360997
             EXP
 CAL   0.029406601
LODG  -0.009428706
ELEC  -0.043417433
EDUC  -0.055808549
  TV   0.096479783


[[13]] :
              POP         DENS         POPG         AGRF         INDU          GNP
 CAL   -1.2381463   0.32070270  -0.50182691  -1.5559284  -1.46659760  -0.329970465
LODG    1.5844712   0.05936443   0.84575304   1.4093584   0.26007844  -0.451451148
ELEC   -0.2266287  -0.60797688  -0.09479546  -0.8424598  -0.29547841  -0.007662604
EDUC    0.9399734   0.12062866   0.19297576  -0.9464048   0.22764572   0.499109472
  TV    0.7560378  -0.17603215   0.26503669  -0.4111226   0.08547898  -0.036602828
             GDPA          FCF          RR          OFR          DR          IMP
 CAL   0.28156306  -0.29973321   0.03477639   2.2097192  0.110538640   4.5577070
LODG  -0.71687517   0.07836553   1.20967477  -2.0487535  0.507546127  -5.2355379
ELEC  -0.09813817   0.32452517   0.13135762  -0.1191081 -0.421821919   0.1780810
EDUC   1.20093406   0.33318454   0.82032933  -0.1994464  0.008561542  -0.3738255
  TV   0.06410141  -0.14344311   0.37649124  -0.7592389  0.254624064  -2.3818107
             EXP
 CAL   -5.2584918
LODG    5.8968501
ELEC    0.1475560
EDUC   -0.4767778
  TV    2.7663124
```

Les données [[**A**]] représentent les coefficients du modèle de régression linéaire

$$\widehat{Y}_A = X\widehat{\beta}_A$$

en dimension A, c'est-à-dire $[[A]] = \widehat{\beta}'_A$ qui est une matrice 5×13.

On peut donc lire, en dimension 6 :

$\widehat{CAL}_6 = -0.434POP + 0.455DENS + 0.179POPG - 0.152AGRF - 0.492INDU + 0.099GNP +$

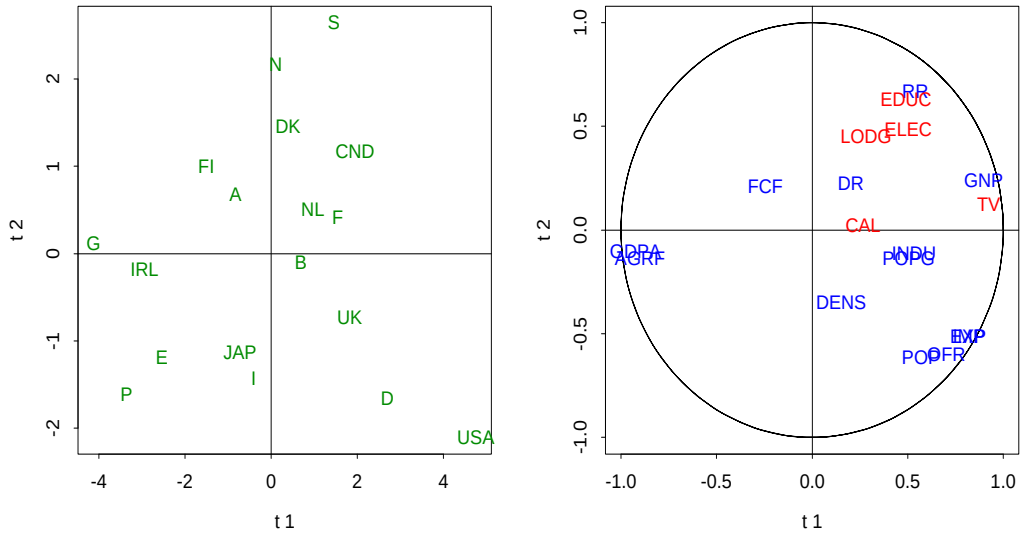
$0.283GDPA - 0.766FCF + 0.236RR + 0.248OFR + 0.329DR + 0.085IMP + 0.029EXP$

On note que le modèle PLS ($A = 6$) présente plus de cohérence que le modèle OLS

$(A = 13)$: les coefficients des variables IMP et EXP sont de même signe et agissent donc sur CAL dans le même sens, ce qui n'est pas le cas pour OLS.

$\widehat{CAL}_{13} = -1.238POP + 0.320DENS - 0.501POPG - 1.555AGRF - 1.466INDU - 0.329GNP + 0.281GDPA - 0.299FCF + 0.034RR + 2.209OFR + 0.110DR + 4.557IMP - 5.258EXP$

```
 cercle des correlations pour les t avec X et Y (o/n)?
1 :  o
```



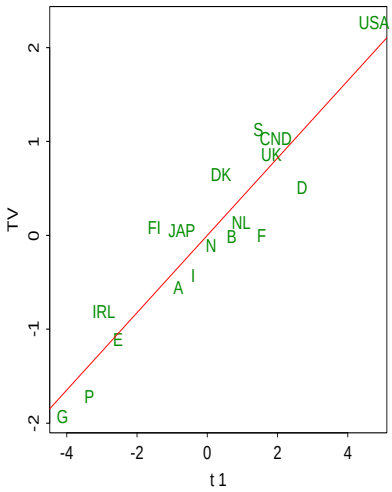
La fonction **pls** permet d'afficher côte à côte le cercle des corrélations dans un plan (t_i, t_j) et le graphe des projections des points-individus sur ce même plan.

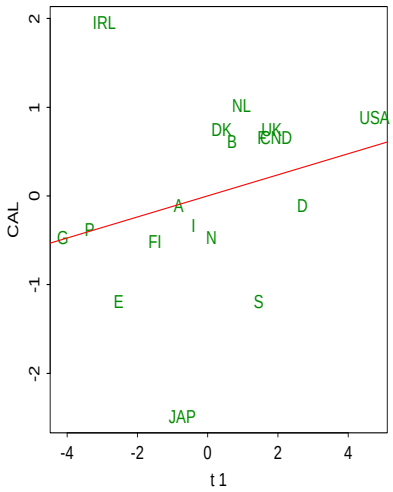
Cet affichage est plus pratique pour l'interprétation simultanée permettant de "donner un nom" aux composantes $t_1, ..., t_A$. Le graphe des projections des points-individus dans le plan (t_i, t_j), avec $i, j \leq A$, où A est le nombre total de composantes choisi pour le modèle, permet en fait de visualiser les individus dans des plans résumant au mieux les prédicteurs, tout en étant orientés vers l'explication des réponses.

Les deux graphiques précédents sont pratiquement identiques à ceux de l'ACP sur le tableau `ocde` des mesures sur toutes les variables tant explicatives que réponses. Remarquons seulement une anomalie dans les graphes de l'ACP et de PLS : les USA semblent dépenser pour l'éducation publique (EDUC) moins que la moyenne des pays

considérés. Si l'on retourne aux moyennes des variables, il n'en est rien, 5.1 pour EDUC des USA alors que la moyenne $\overline{\text{EDUC}} = 4.73$. Cela semblerait signifier que les USA ne dépensent pas assez pour l'éducation publique comparativement à leur niveau de richesse. En d'autres termes, la relation entre EDUC et les autres variables n'est peut-être pas linéaire. L'utilisation des splines ou polynômes par morceaux pour transformer les prédicteurs de la régression PLS mettra en évidence le déficit de participation des USA à leur éducation publique compte tenu de leur potentiel économique.

```
 regression des Y par les t (o/n)?
1 :  o
```





Ces graphiques prennent leur sens dans le fait que le modèle PLS s'écrit comme une somme de modèles linéaires partiels en les composantes principales t_k

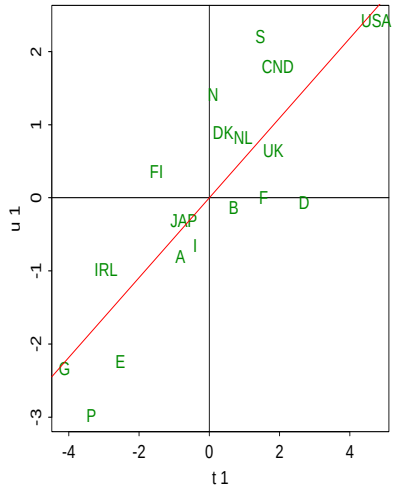
$$\widehat{Y}_A = \sum_{k=1}^{A} P_{t_k} Y \ .$$

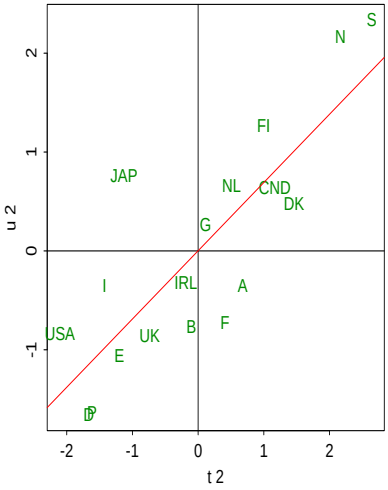
La matrice P_{t_k} est la matrice $n \times n$ de projection orthogonale sur la composante t_k.
Sur le graphe précédent, TV est très bien expliquée par t_1. Si l'on devait seulement faire la régression de TV, on choisirait A égal à 1.
De même, la régression de CAL sur t_1 est mauvaise, ce qui est normal car elle est

seulement expliquée à 7.1 % par t_1.

```
 regressions des u par les t (o/n)?
```

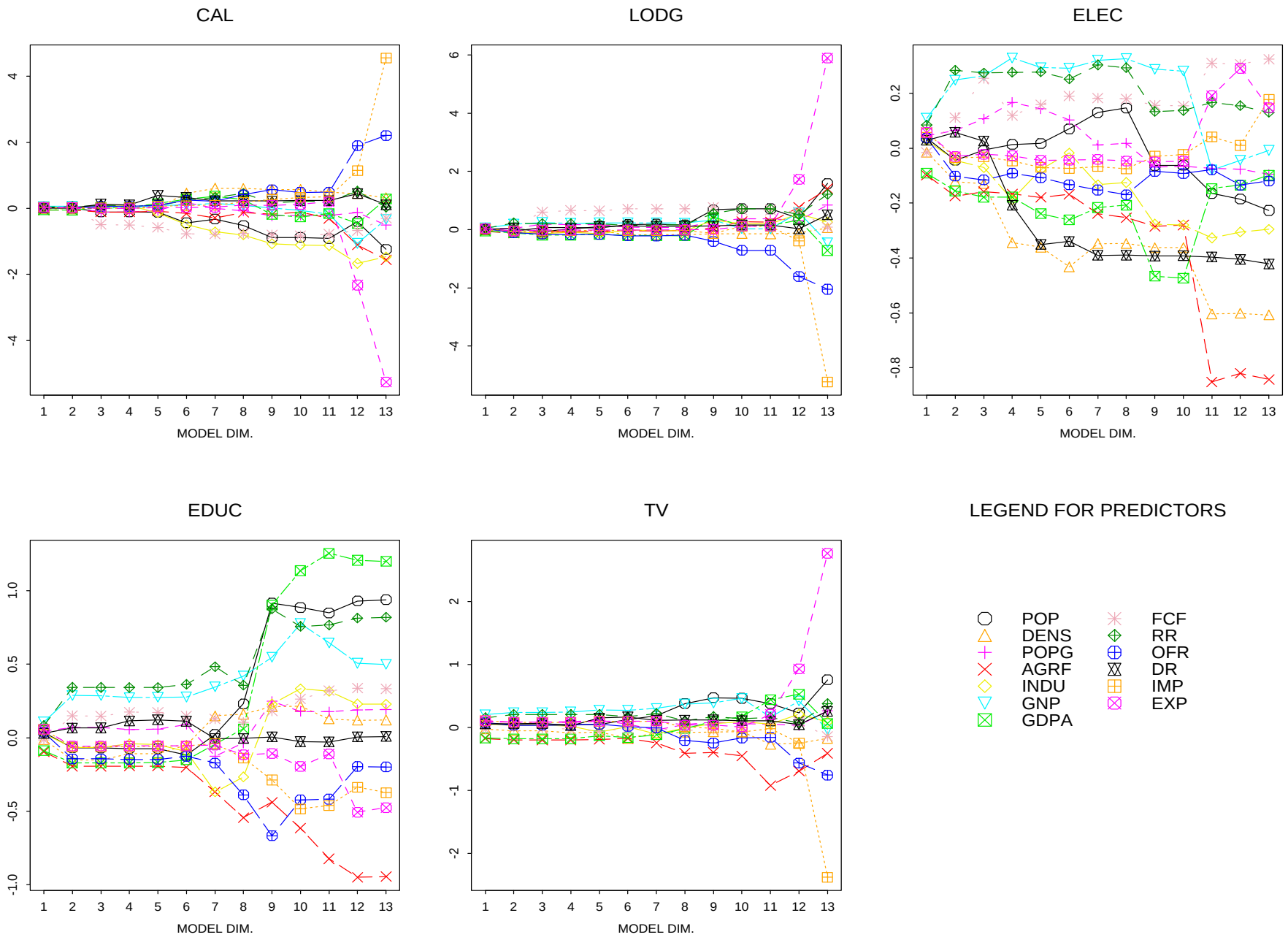
1 : o





Dans le cas d'une régression PLS multiréponses, les graphiques (t_k, u_k) sont des indicateurs de qualité pour les régressions linéaires partielles aux différentes étapes k de PLS. En effet, les u peuvent être considérés comme des composantes principales des réponses.

```
 evolution des coeffs des modeles suivant le nombre de composantes
(o/n)?
1 :  o
un plot pour Chaque reponse ou un plot pour Toutes les reponses (C/T)
1 :  T
Nombre de graphes par ligne ?  (<= 6 , prevoir 1 plot de plus pour la
legende)
1 :  3
Nombre de lignes ?
1 :  2
Click on the plot to locate the legend !
```


CAL
LODG
ELEC
EDUC
TV
LEGEND FOR PREDICTORS
MODEL DIM.
POP
DENS
POPG
AGRF
INDU
GNP
GDPA
FCF
RR
OFR
DR
IMP
EXP

Ces graphes permettent de corroborer le choix du nombre A de composantes principales déjà effectué lors du déroulement pas à pas de l'algorithme. C'est un critère d'ajustement basé sur l'étude de l'évolution des coefficients des modèles $\widehat{Y}_A = X\widehat{\beta}_A$ dimension par dimension.

Pour une réponse j choisie, l' étendue des valeurs du vecteur $\widehat{\beta_A^j}$ reste stable jusqu'à une certaine valeur de A. Au-delà de cette valeur de A, pour certains prédicteurs fortement corrélés, les valeurs des coefficients $\widehat{\beta_A^j}$ semblent exploser. On choisit alors la plus petite valeur de A à partir de laquelle apparaît ce phénomène.

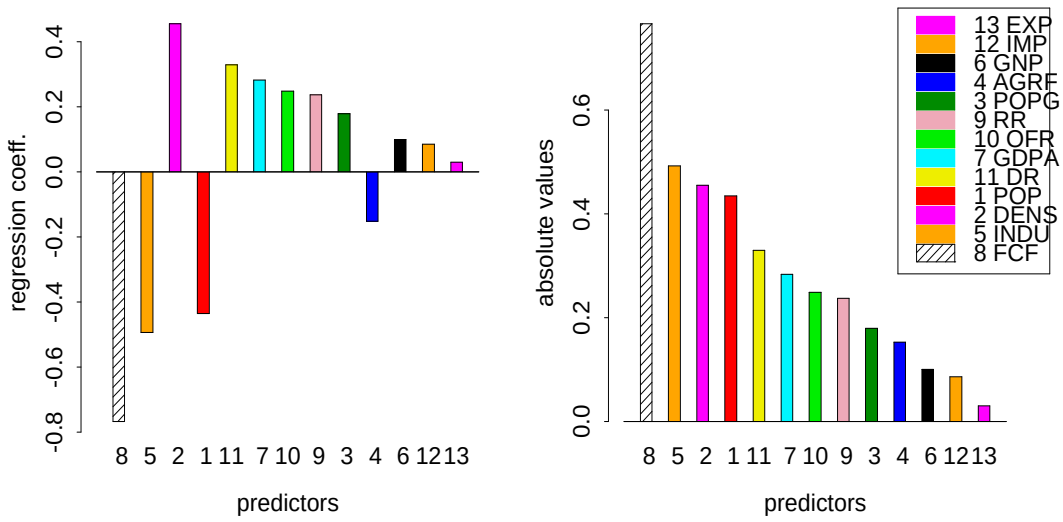
Ici, l'intervalle de stabilité des coefficients des variables CAL et LODG s'arrête en dimension 11, celui de EDUC en dimension 6, ceux de ELEC et TV en dimension 10. Par conséquent, ce critère nous amène à retenir le modèle :

$$\widehat{Y}_6 = X\widehat{\beta}_6.$$

Ici, l'explosion des coefficients de EDUC à partir de la dimension 6 est très visible pour les prédicteurs fortement corrélés, AGRF et GDPA d'une part, IMP et EXP d'autre part.

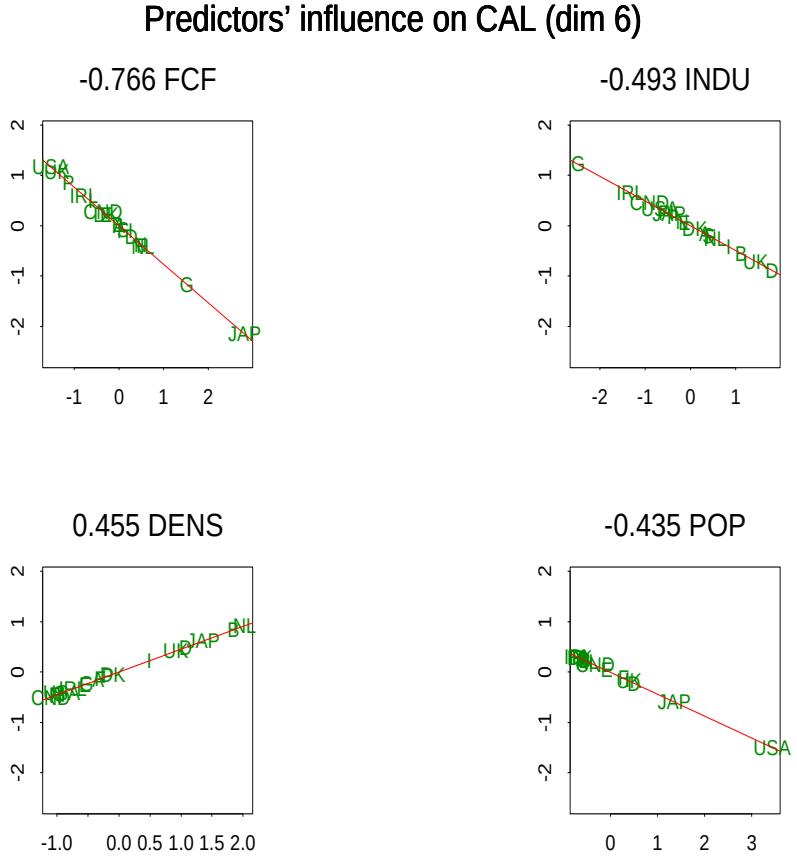
```
 representation des predicteurs de plus grande influence (o/n)?
1 :  o
numero de la reponse (<= 5 ) ?
1 :  1
nb d'axes du modele (<= 13 ) ?
1 :  6

   FCF   INDU   DENS    POP     DR   GDPA    OFR     RR   POPG   AGRF
 0.766  0.493  0.455  0.435  0.329  0.283  0.249  0.237  0.179  0.152
   GNP    IMP    EXP
   0.1  0.085  0.029
```

Les variables explicatives sont classées par ordre décroissant des valeurs absolues de leurs coefficients dans le modèle de dimension 6.



Le premier graphe représente les valeurs des coefficients pour chacun des prédicteurs, et le deuxième leurs valeurs absolues respectives. Pour la variable réponse considérée et dans la dimension choisie, les prédicteurs sont classés par ordre décroissant des valeurs absolues de leurs coefficients. Pour la variable CAL en dimension 6 les prédicteurs les plus influents sont : FCF, INDU, DENS, POP, où FCF, INDU et POP influent négativement.

```
 Combien de predicteurs ?  :
1 :  4
Combien de plots par ligne ?,(<= 4 )
1 :  2
Combien de ligne(s) ?
1 :  2
```



C'est une autre représentation des variables les plus influentes. Le premier graphe indique qu'un taux de formation du capital fixe important pour un pays "abaisse" la consommation de calories par habitant. Le Japon semble se distinguer des autres quant à la variable FCF : il a un taux de formation de capital fixe nettement supérieur aux autres pays. Le graphique concernant la population (POP) indique que, d'après ce modèle ($\widehat{CAL}_6 \simeq -0.766FCF - 0.493INDU + 0.455DENS - 0.435POP$), plus un pays est peuplé, plus la consommation de calories par habitant diminue. En regardant la position des USA sur les quatre graphes, il semblerait que les USA aient une consommation de calories par habitant un peu inférieure à la moyenne. Or, d'après les données ils ont une des valeurs les plus fortes pour cette variable. Les USA se distinguant énormément des autres pays pour la variable POP, leur valeur pour ce prédicteur "annule" les effets des trois autres variables influentes, ce qui fausse la prédiction pour ce pays. Pour remédier à ce problème, il existe deux autres méthodes de modélisation non linéaire par PLS, que nous ne développerons pas ici, qui donnent de meilleurs ajustements avec moins de composantes : la régression PLS linéaire sur codage spline des prédicteurs ou PLSS

(Durand 1997) et la régression PLS non linéaire sur splines additives ou ASPLS (Durand et Sabatier 1997).

Ces méthodes modèlisent de façon non linéaire additive les réponses en fonction des prédicteurs

$$\widehat{y}_A^j = \sum_{i=1}^{p} f_i^j(x_i) \ ,$$

où la *fonction coordonnée* f_i^j s'interprète comme l'influence du prédicteur x_i sur la réponse y_j. Ces fonctions coordonnées dépendent comme dans le modèle linéaire, du nombre A de composantes PLS utilisées.

```
 plot des erreurs des reponses suivant le nb de composantes (o/n)?
1 :  o
Combien de composantes ?,(<= 13 )
1 :  7
```

Ces plots nous permettent de visualiser l'importance de l'erreur faite sur CAL, LODG, ELEC, EDUC et TV lors de leur estimation dans chaque dimension : plus l'erreur est proche de zéro, mieux sont estimées les réponses. Sur les plots ci-après, l'erreur pour chaque pays se lit sur l'axe des ordonnées. Sur l'axe des abscisses sont portées les valeurs des réponses estimées.

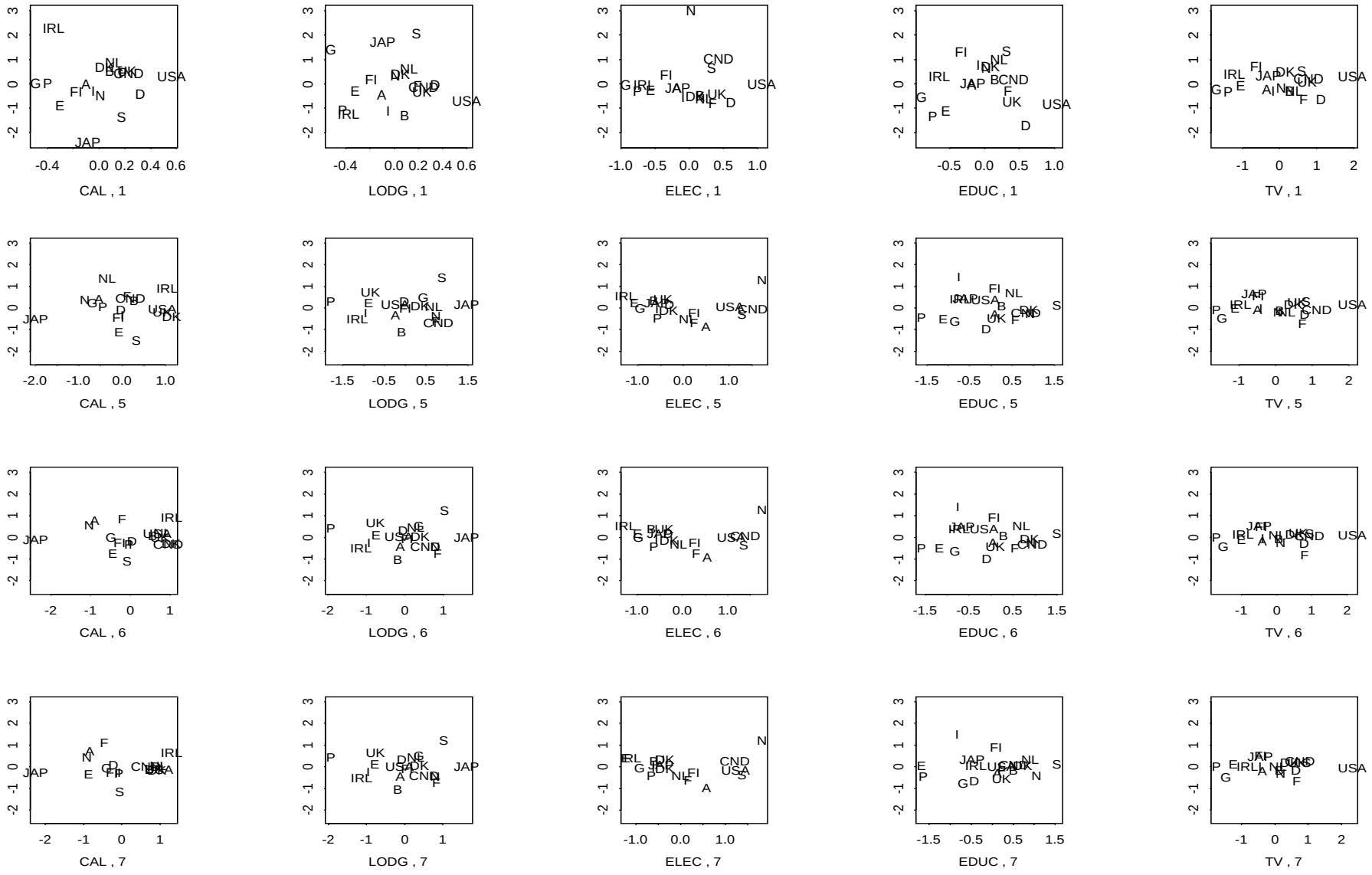
On peut remarquer que l'évolution des erreurs entre la dimension 5 et la dimension 6 n'est pas flagrante, l'erreur semble se stabiliser.

De plus, on voit bien qu'en dimension 1, l'erreur sur la variable TV est très petite : on retrouve le fait que le R^2part entre TV et la composante 1 est de 0.85.

Remarque : Ce graphique peut servir d'aide au choix de A. Il permet de visualiser à partir de quelle dimension les erreurs sur les réponses ne diminuent plus de façon significative.

Sur la page ci-après sont présentés seulement les plots des erreurs en dimension 1 puis en dimensions 5, 6 et 7.

Errors according to PLS dimensions



3.3 Validation du nombre de composantes et prédiction

Le choix du nombre d'axes intervient lors du déroulement de l'algorithme par un critère basé sur la qualité de l'ajustement aux données (pourcentage de variance expliquée pour les réponses et les variables explicatives). Ensuite, l'examen des coefficients des modèles calculés pour des dimensions différentes permet de mettre en évidence des zones de stabilité qui, en quelque sorte, confirment le choix précédent. Ces deux critères sont cependant insuffisants pour être rassuré sur la qualité de prédiction du modèle PLS. Deux critères supplémentaires sont utilisés pour valider cette dernière. Le premier, critère de validation interne, est basé sur la validation croisée. Le second, critère de validation externe, nécessite la connaissance d'un jeu de données supplémentaire appelé échantillon à tester. Lorsque l'échantillon test ne contient que les variables explicatives sans les réponses, il n'est plus question de valider mais de prédire. C'est ce que fait la fonction **pls**.

3.3.1 Critère de validation croisée

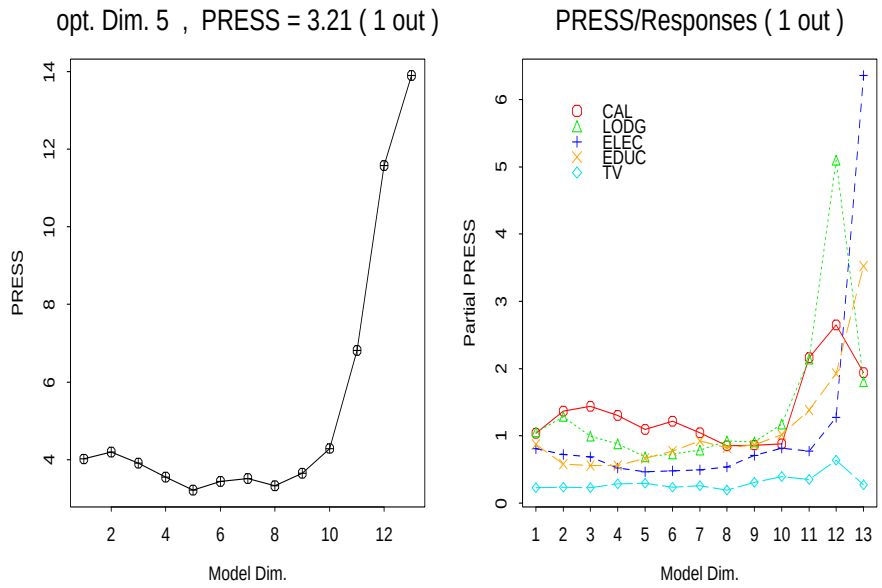
Après avoir enlevé un individu i (une ligne) aux matrices X et Y (ou 10% des individus si leur nombre est élevé), on calcule $\widehat{\beta}_A^{(i)}$, qui représente la matrice des coefficients du modèle construit avec les individus restants. Puis, on calcule l'erreur de prédiction faite sur l'individu i, $E_A^{(i)} = Y_i - X_i\widehat{\beta_A}^{(i)}$. La qualité de prédiction est définie par le PRESS, PREdiction Sum of Squares :

$$PRESS(A) = \frac{1}{n}\sum_{i=1}^{n} \|E_A^{(i)}\|^2.$$

On choisit donc A tel que $PRESS(A)$ soit le plus petit possible.

Les deux fonctions **plscv** et **plscv.plot** permettent de calculer et de visualiser l'évolution des valeurs de PRESS(A) :

```
>ocdeplscv<-plscv (ocdeX,ocdeY,A=13,prop=0.05)
>plscv.plot (ocdeplscv)
```



Le premier graphe représente le PRESS total qui est la somme des PRESS partiels pour chaque réponse représentés sur la deuxième figure. La dimension optimale est 5.

3.3.2 Critère de validation externe et prédiction

Pour utiliser ce critère de choix de A, on doit posséder outre l'échantillon d'apprentissage (X, Y) sur les n individus, un échantillon test $(Xtest, Ytest)$ sur N observations mesurées sur les mêmes variables. Différents modèles PLS basés sur (X, Y) sont calculés pour différentes dimensions. La valeur optimale de A correspond à la plus petite des erreurs de prédiction calculée sur $(Xtest, Ytest)$. L'exemple présenté a pour seul intérêt de montrer comment utiliser ce critère sous Splus. En effet, l'échantillon test est basé sur un nombre d'observations nettement insuffisant pour espérer valider la dimension du modèle.
L'échantillon d'apprentissage (ocdeXa, ocdeYa) proposé est construit à partir de (ocdeX, ocdeY) sans les quatre derniers pays. L'échantillon test (ocdeXt, ocdeYt) est formé de ces quatre derniers pays.

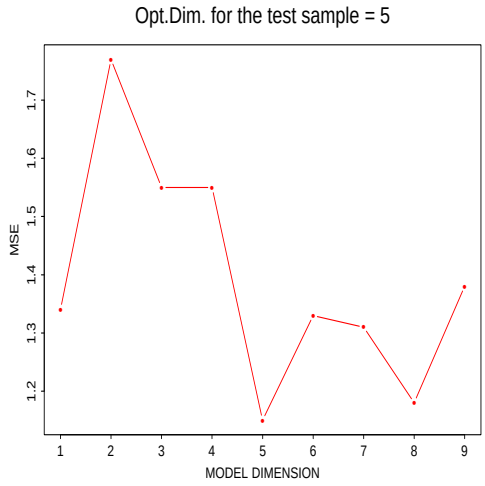
```
>ocdeXa<-ocdeX[-(15 :18),]
>ocdeYa<-ocdeY[-(15 :18),]
>ocdeXt<-ocdeX[15 :18,]
>ocdeYt<-ocdeY[15 :18,]
>pls(ocdeXa,ocdeYa,ocdeXt,ocdeYt)
```

Les premières sorties numériques correspondent à celles du programme **pls** usuel, dans lesquelles se trouvent les coefficients $\widehat{\beta}_A$.

Ensuite apparaît :

```
Moyenne des carres des erreurs (MSE) sur variables reduites suivant la
dimension du modele
       1    2    3    4    5    6    7    8    9
MSE 1.34 1.77 1.55 1.55 1.15 1.33 1.31 1.18 1.38
Prediction lineaire des reponses a partir de l'echantillon test ?(o/n)
Choisissez la dimension du modele (<= 9 )
```

MSE représente la moyenne des carrés erreurs de prédiction sur l'échantillon test. Ici on voit qu'en dimension 5, l'erreur est la plus petite. On choisirait donc A=5 composantes principales pour le modèle PLS. Le graphique ci-après illustre ce choix.



A la question posée, on répond donc 5 pour obtenir les erreurs sur chaque variables de la matrice test des réponses.

```
Reponses estimees et erreurs de prediction  Yerr = Ytest - Yest
    est.CAL    err.CAL est.LODG  err.LODG est.ELEC   err.ELEC est.EDUC
NL 2876.300  363.69954 8.532376 1.1676244 5117.483 -2552.4828 5.424635
 P 2796.047  133.95304 4.171352 0.1286479 3387.522 -2780.5218 3.217873
UK 3227.531  -47.53141 4.664680 3.0353200 2470.410  1209.5904 5.205442
 S 3251.251 -501.25092 8.031398 5.3686020 7552.773  -749.7726 6.741364
     err.EDUC    est.TV     err.TV
NL  1.2753652 199.67286  -2.672863
 P -1.8178734  43.64337 -14.643371
UK -1.0054425 228.07467  34.925334
 S  0.6586365 247.31494  40.685060
Erreurs moyennes sur les reponses :
  err.CAL err.LODG err.ELEC err.EDUC   err.TV
 317.6997 3.139066 2016.912 1.262531 27.82367
Erreur moyenne globale =  473.367
```

L'erreur moyenne d'une réponse est la racine carrée de la moyenne des carrés des erreurs pour chaque individu.

```
>pls(ocdeXa,ocdeYa,Xtest=ocdeXt)
```

Cette instruction permet de prédire les réponses à partir de l'échantillon $Xtest$. Il est à noter que lorsque $Xtest$ est présent, les individus correspondants sont représentés en individus supplémentaires dans les graphiques des composantes.

3.4 PLS considérée comme une ACP

On va vérifier numériquement que PLS(X,X)=ACP(X). Pour cela, on va voir que les composantes principales de ocdeX calculées lors de l'ACP de X sont les mêmes que les composantes principales calculées lors de PLS(X,X).

```
>composantespls<-pls(ocdeX,ocdeX,A=4)$TX
>composantesacp<-acpxqd(ocdeX,k=4)$C
```

	c1	c2	c3	c4	t1	t2	t3	t4
D	3.695	-0.068	-1.506	0.331	3.695	-0.068	1.506	0.331
A	-1.179	-0.975	0.117	0.845	-1.179	-0.975	-0.117	0.845
B	0.754	-1.911	-1.569	0.138	0.754	-1.911	1.569	0.138
CND	1.522	0.775	2.055	-0.998	1.522	0.775	-2.055	-0.998
DK	-0.460	-1.820	0.683	-1.080	-0.460	-1.820	-0.683	-1.080
E	-2.159	1.254	-0.346	0.452	-2.159	1.254	0.346	0.452
USA	5.841	3.247	-1.716	0.543	5.841	3.247	-1.716	0.543
FI	-2.082	-0.147	0.920	-0.055	-2.082	-0.147	-0.920	-0.055
F	1.467	-0.278	0.203	-0.981	1.467	-0.278	-0.203	-0.981
G	-4.196	2.087	-0.121	-1.479	-4.196	2.087	0.121	-1.479
IRL	-3.267	0.218	0.593	0.527	-3.267	0.218	-0.593	0.527
I	0.198	0.349	-0.784	1.329	0.198	0.349	0.784	1.329
JAP	0.279	1.996	-2.662	-1.939	0.279	1.996	2.662	-1.939
N	-0.806	-1.055	1.291	-0.007	-0.806	-1.055	-1.291	-0.007
NL	0.966	-1.317	-1.296	-0.515	0.966	-1.317	1.296	-0.515
P	-2.826	1.602	-0.212	2.255	-2.826	1.602	0.212	2.255
UK	2.043	-1.645	-0.683	1.282	2.043	-1.645	0.683	1.282
S	0.208	-2.311	1.601	-0.647	0.208	-2.311	-1.601	-0.647

Cette propriété permet de considérer l'ACP usuelle comme une méthode de régression PLS. Ainsi, l'introduction dans PLS de métriques sur l'espace des individus permettrait de considérer les différentes méthodes d'Analyse de Données comme des régressions PLS particulières.

ANNEXE

Annexe 1 : pour saisir les données sous Splus...

L'objet du type vecteur est très important sous Splus. Un objet d'un seul élément sera considéré comme un vecteur de longueur 1. C'est la fonction **c** qui crée un vecteur dont la longueur est le nombre d'éléments donnés :

```
>vect<-c(1,2,3)
```

On a ici créé un vecteur de composantes : 1, 2, 3.

Remarque : <- est une flèche d'affectation, qui peut tout aussi bien se noter _. Sous Splus, en général, on écrit d'abord le nom de l'objet créé, puis la flèche d'affectation et enfin, la fonction qui permet de créer l'objet en question. Dans le cas de l'exemple, vect est ainsi le nom du vecteur que l'on a créé. Si on veut obtenir le contenu de ce vecteur, et plus généralement le contenu d'un objet de Splus, il suffit de taper le nom de l'objet :

```
>vect
```

Pour créer un vecteur, il est également possible d'utiliser la fonction **scan** :

```
>vect<-scan()
```

L'ordinateur retourne

```
1 :
```

Il reste alors à taper les valeurs du vecteur (1, 2, 3 si on veut obtenir le même vecteur qu'auparavant) :

```
>1 :  1 2 3
```

Il suffit ensuite, de faire la même manipulation que précédemment, pour obtenir le contenu du vecteur.
D'autre part, la fonction **scan** offre aussi la possibilité de lire un vecteur situé dans un fichier. Supposons l'existence d'un fichier qui contienne uniquement ce vecteur,

```
>vectbis<-scan("vectfich")
```

signifie que le vecteur situé dans le fichier vectfich est devenu le vecteur vectbis sous Splus.
Pour créer une matrice, c'est la fonction **matrix** qu'il faut utiliser.

```
>X<-matrix(c(2,5,3,1,4,2),ncol=3,byrow=T)
```

Dans l'exemple ci-dessus, X est le nom que l'on a donné à la matrice, puis, comme arguments, on trouve respectivement, le vecteur des valeurs, le nombre de colonnes que comporte la matrice, et enfin, byrow=T signifie que matrix( ) lit les valeurs ligne par ligne.
Ici, il y aura donc 2 lignes.
Par défaut (si l'on omet byrow=T), la matrice sera lue colonne par colonne.

```
>X[2,1]
```

permet d'obtenir l'élément situé à l'intersection de la deuxième ligne et de la première colonne (1 dans le cas de l'exemple).
On peut aussi sélectionner des lignes et des colonnes d'une matrice. Les commandes qui vont suivre permettent d'obtenir respectivement la troisième ligne et la deuxième colonne de la matrice X :

```
>X[3,]
```

```
>X[,2]
```

De plus, on peut changer le nom des lignes et des colonnes à l'aide de la fonction **dimnames** :

```
>dimnames(X)<-list(c(''l1'',''l2''),c(''r1'',''r2'',''r3''))
```

Celle-ci fait intervenir une autre fonction de Splus : **list**. Cette dernière admet comme argument des objets qui peuvent être de divers types; dans le cas ci-dessus, ce sont des vecteurs qui représentent respectivement, les nouveaux noms des 2 lignes et des 3 colonnes.

Par ailleurs, Splus donne la possibilité de créer et de manipuler un autre type d'objets : les tableaux de données (data frame en anglais). Ce sont des généralisations des matrices qui présentent 2 avantages comparés aux matrices classiques :
_ ils permettent d'utiliser des valeurs numériques et des chaînes de caractères...
_ ils attribuent des noms aux lignes et aux colonnes (plus simplement que pour les matrices classiques). De plus, ces nouveaux noms pourront être utilisés pour visualiser le contenu d'une variable, ce qu'il est impossible de faire avec une matrice classique même si l'on a pris soin de changer le nom des lignes et des colonnes avec la fonction **dimnames**.

Tout d'abord, écrivons dans un éditeur de texte le tableau de données suivant que l'on sauve sous le nom notes :

	Math.	Phys.	Chim.
Arthur	3	Abs	9
Jules	10	9	12
Sophie	Abs	13	11
Arlette	Abs	Abs	14

Ensuite, sous Splus

```
>Xbis<-read.table(‘‘notes’’,header=T)
```

Ainsi Xbis est un data frame. Pour accéder à une des variables de la matrice, on utilise la fonction **attach**.

```
>attach(Xbis)
```

Cette commande permet d'identifier les colonnes du tableau par leur nom, par exemple

```
>Math.
```

```
>detach()
```

Une commande très utile permet de transformer des matrices en data frame : **as.data.frame**

```
>Xtransform<-as.data.frame(X)
```

Si l'on veut revenir à l'objet matriciel,

```
>Xtransform<-as.matrix(Xtransform)
```

Annexe 2 : Description de fonctions Splus utilisées

Dans cette annexe, sont présentées les fonctions non natives à Splus utiles pour faire l'analyse de données. Pour avoir des détails sur les fonctions natives à Splus, il suffit d'utiliser les fonctions **help()** ou **help.start(gui="motif()")**.

↪ **Dvar**

Cette fonction retourne une liste. Tous les objets de cette liste sont précédés d'un $.

entrées	*description*	*valeurs par défaut*
X	matrice	
D="vector"	vecteurs des poids des individus; équirépartis si =1	1
cor="bool"	si =F, D-centrage de X si =T, D-centrage et réduction de X	F

Les sorties :

V : matrice des D-variances de X si cor=F, des D-corrélations si cor=T,

U : matrice D-centrée de X si cor=F, D-centrée-réduite de X si cor=T,

mean : vecteur des D-moyennes des colonnes de X,

var : vecteur des D-variances des colonnes de X.

↪ **Dproj**

Cette fonction permet de projeter les réponses Y sur les prédicteurs X au sens de la métrique D.

entrées	*description*	*val. par défaut*
X	matrice des prédicteurs	
Y	matrice des réponses	
D	vecteur des poids sur les individus	1
eps	scalaire seuil du zéro numérique	$1e^{-08}$

Les sorties :

$H : matrice de la projection $H = X(X'DX)^{+}X'D$.

$Yhat : projection de Y sur l'e.v. engendré par les X, $Yhat = HY$.

$Yres : matrice des erreurs, $Yres = Y - Yhat$.

$beta : coefficients de la régression, $\beta = (X'DX)^{+}X'DY$.

↪ **acpxqd**

Cette fonction permet de faire une ACP généralisée ou une ACP réduite.

entrées	*description*	*val. par défaut*
X	matrice des variables	
Q	métrique des individus (u.s.); si Q=1 alors Q=Idn si Q est un vecteur alors Q=diag(vecteur), sinon Q=matrice	1
D	métrique des variables; si D=1 alors D=1/nIdn si D est un vecteur alors D=diag(vecteur), sinon D=matrice	1
centrer="bool"	si =F on ne centre pas X si =T on centre X	T
cor="bool"	si =F on ne réduit pas si =T centrage et réduction	T
k="num"	nombre de composantes retenues	3
impres="bool"	impression des résultats si T si F rien	T
graph="bool"	trace des graphiques si T, rien si F	T
Xl="matrix"	matrice éventuelle des individus supplémentaires	
Xc="matrix"	matrice éventuelle des variables supplémentaires	
aideus	si 1 aides à l'interprétation pour les u.s. CTR et COS pour les k premiers axes; si =0 rien	1
aideva	idem pour les variables	1
cexpar	valeur du paramètre cex pour les plots sous Splus	1
colpar="entier"	numéro de la couleur pour les plots	1

↪ **plscv**

Cette fonction sert à calculer le PRESS du critère de validation croisée (cf. Section 2.3.1).

entrées	*description*	*valeurs par défaut*
X	matrice des variables explicatives	
Y	matrice des réponses	
StandX="bool"	si =T, X est D-centrée-réduite; si =F D-centrée seulement	T
StandY="bool"	idem	T
A="num"	nombre de composantes de la PLS	2
D="vector"	vecteur des poids équirépartis	1
prop	proportion d'individus prédits par les individus restants	0.1

Cette fonction n'a pas de sorties graphiques : il faut l'utiliser comme argument de la fonction qui suit :

↪ **plscv.plot**

Cette fonction donne le graphique du PRESS.

entrées	*description*	*valeurs par défaut*
plscvresul	c'est l'objet Splus obtenu par la fonction plscv	
cexpar	donne la grandeur des caractères de la légende, cexpar+0.5 est le paramètre cex pour les plots	1
ncolpar	entier donnant le nombre de colonne de la légende	1
colpar	numéro de la couleur pour les plots	1
titlepar	booléen, si T donne un titre aux plots	T

Le premier plot donne le PRESS total (somme des PRESS partiels pour chaque réponse).

Le deuxième donne les PRESS partiels pour chaque réponse.

Pour localiser le coin supérieur gauche de la légende, cliquer sur le plot.

↪ **pls**

entrées	*description*	*val. par défaut*
X	matrice des var. explicatives	
Y	matrice des réponses	
Xtest	matrice des prédicteurs de l'ech.test s'il y a	
Ytest	matrice des réponses de l'ech. test	
StandX="bool"	si =T, X est D-centrée-réduite; si =F, D-centrée seulement	T
StandY="bool"	idem	T
D="vector"	vecteur des poids équirépartis	1
A="num"	nombre de composantes de la PLS	1
eps="num"	scalaire seuil du zéro numérique pour les valeurs singulières (calcul du projecteur)	$1e^{-08}$
splflag="bool"	si =F, PLS linéaire; si =T, PLS sur codage spline des prédicteurs	F
impres="bool"	T⇒ impression des résultats; F⇒ rien	T
graph="bool"	T⇒ trace des graphiques; F ⇒ pas de trace	T
typedata ="bool"	si T, le plot des résidus donne le nom des observations si F, leur numéro seulement	T
ptypar, cexpar	valeurs des paramètres pty et cex pour les plots sous Splus	"s" ;1
titlepar="bool"	si T, met un titre sur les plots multiples	T
colpar="entier"	numéro de la couleur pour les plots	1

Annexe 3 : Un peu de Mathématique PLS

La méthode PLS est une méthode de régression linéaire de q variables réponses sur p variables explicatives toutes mesurées sur les mêmes n individus. Les tableaux des observations, notés respectivement Y et X, de dimensions $n \times q$ et $n \times p$, sont supposés centrés et éventuellement réduits par rapport aux poids $(p_1, \ldots, p_n)$. On note $D = diag(p_1, \ldots, p_n)$ la matrice diagonale des poids.

L'intérêt de la méthode comparée à la régression sur composantes principales (RCP), voir la Section 3.1, réside dans le fait que les composantes PLS sur les X, notées t, sont calculées "dans le même temps" que des régressions partielles sont exécutées. Cette simultanéité leur confère un meilleur pouvoir prédictif que celles de la RCP. La question est donc d'examiner comment cette simultanéité est mise en oeuvre.

Notons $E_0 = X$ et $F_0 = Y$ les tableaux centrés et réduits au sens de D qui en général est égal à $n^{-1} I_n$. La méthode procède par étapes successives permettant le calcul des composantes principales. On notera A le nombre total d'étapes, c'est-à-dire de composantes indicées par $k = 1, \ldots, A$.

Description de la k-ième étape

Notons $t = E_{k-1} w$ et $u = F_{k-1} c$, les combinaisons linéaires colonnes des matrices centrées E_{k-1} et F_{k-1} associées respectivement aux vecteurs des poids w et c. La covariance entre t et u s'écrit comme le D-produit scalaire

$$cov(t, u) = (t, u)_D = w' E'_{k-1} D F_{k-1} c \ .$$

Le carré de la D-norme associée fournit la variance $\|t\|^2_D = var(t)$.

L'étape k, se décompose en deux parties. La première fournit les composantes $t_k = E_{k-1} w_k$ et $u_k = F_{k-1} c_k$ par le calcul des poids optimaux w_k et c_k. La deuxième actualise les matrices des prédicteurs et des réponses E_k et F_k comme résidus de la régression sur t_k.

Calcul des poids	$(w_k, c_k) = \arg\max cov(t, u) = w' E'_{k-1} D F_{k-1} c$ sous les contraintes $\|w\|^2 = \|c\|^2 = 1$,
Actualisation	$E_k = E_{k-1} - P_{t_k} E_{k-1}$ $F_k = F_{k-1} - P_{t_k} F_{k-1}$,

où $P_{t_k} = t_k t'_k D / var(t_k)$ est la matrice $n \times n$ de projection D-orthogonale sur t_k. Remarquons que le critère à optimiser, la covariance, est un compromis entre le critère de

l'Analyse des Corrélations Canoniques, la corrélation, et celui de l'ACP sur chacun des tableaux, la racine carrée de la variance.

Calcul des composantes t_k

Pour résoudre ce problème on utilise la méthode des multiplicateurs de Lagrange basée sur le Lagrangien

$$L = (t,u)_D + \frac{\lambda}{2}(1 - \|c\|^2) + \frac{\mu}{2}(1 - \|w\|^2)$$

où λ et μ sont les multiplicateurs associés aux contraintes. La résolution des équations normales donne

$$\lambda = \mu = cov(t,u)\,,$$

ce qui conduit à

$$F'_{k-1}DE_{k-1}E'_{k-1}DF_{k-1}c = \lambda^2 c$$

$$E'_{k-1}DF_{k-1}F'_{k-1}DE_{k-1}w = \lambda^2 w\,.$$

La solution (w_k, c_k) est fournie par les vecteurs associés à la plus grande valeur singulière λ_k pour la matrice $E'_{k-1}DF_{k-1}$.

Propriétés des composantes $t_1, \ldots, t_A$

Les formules d'actualisation des variables conduisent à la relation

$$(t_k, t_l)_D = (t_k, u_l)_D = 0\ , \qquad \forall l > k.$$

La non corrélation ou D-orthogonalité, mutuelle entre les composantes $t_1, \ldots, t_A$ a de multiples conséquences. On montre ainsi par récurrence que t_k appartient à $Im\,X$ espace vectoriel engendré par les prédicteurs. Plus précisément, $t_k = X\alpha_k$ avec

$$\alpha_1 = w_1$$

$$\alpha_k = \left[I_p - \sum_{j=1}^{k-1} \frac{\alpha_j \alpha'_j}{\|t_j\|^2_D} X'DX \right] w_k\,, \quad \forall\ k > 1\,.$$

La non corrélation implique en outre que $\sum_{k=1}^{A} P_{t_k} = P_{T_A}$, où $P_{T_A} = T_A(T'_A D T_A)^{-1} T'_A D$ est le projecteur orthogonal sur la matrice $T_A = [t_1|\ldots|t_A]$.

Les deux derniers résultats permettent de considérer P_{T_A} comme le projecteur sur le sous espace de $Im\,X$ engendré par les composantes $t_1, \ldots, t_A$. Dans le cas particulier où $A = rang(X)$, $P_{T_A} = P_X$.

Enfin, dernière conséquence de la non corrélation, la décomposition des variances totales

des réponses et des prédicteurs fournit deux critères pour le choix du nombre A de composantes.

Le modèle PLS

Les formules d'actualisation entraînent l'écriture des modèles linéaires :

$$X = E_0 = \sum_{k=1}^{A} \widehat{X}_k + E_A \doteq \widehat{X}_A + E_A$$

$$Y = F_0 = \sum_{k=1}^{A} \widehat{Y}_k + F_A \doteq \widehat{Y}_A + F_A \,,$$

où $\widehat{X}_k = P_{t_k} E_{k-1}$ et $\widehat{Y}_k = P_{t_k} F_{k-1}$ sont les modèles partiels de rang 1. $\widehat{X}_A$ est l'approximation de X avec une erreur E_A, idem pour $\widehat{Y}_A$.

L'actualisation des variables et la non corrélation des composantes conduisent à écrire plus simplement les modèles partiels : $\widehat{X}_k = P_{t_k} X$ et $\widehat{Y}_k = P_{t_k} Y$. La non corrélation des composantes permet d'une part la décomposition de la variance totale

$$var(Y) = \sum_{j=1}^{q} var(Y^j) = \sum_{k=1}^{A} var(\widehat{Y}_k) + var(F_A)$$

pour ce qui concerne les réponses. D'autre part, elle conduit à l'écriture définitive des modèles PLS en fonction des composantes

$$\widehat{Y}_A = P_{T_A} Y$$

$$\widehat{X}_A = P_{T_A} X \,.$$

Le projecteur s'écrit aussi

$$P_{T_A} = \sum_{k=1}^{A} \frac{X \alpha_k \alpha_k' X' D}{\|t_k\|_D^2}$$

ce qui implique que le modèle PLS est linéaire en les variables explicatives initiales

$$\widehat{Y}_A = X \widehat{\beta}_A$$

avec

$$\widehat{\beta}_A = \sum_{k=1}^{A} \frac{\alpha_k \alpha_k'}{\|t_k\|_D^2} X' D Y \,.$$

Cas particuliers

- Si A=rang(X), PLS(X,Y)=OLS(X,Y).

  Lorsque A=rang(X), $t_1, ..., t_A$ forment une base D-orthogonale de ImX. Alors, $\widehat{X}_A = X$ et $E_A = 0$. X est entièrement reconstitué. De plus, $\widehat{Y}_A = P_X Y$, ce qui signifie que la régression PLS linéaire est équivalente à q régressions linéaires multiples aux moindres carrés usuels. En règle générale, la régression PLS multiréponses ne conduit pas aux mêmes modèles que ceux obtenus par q régressions PLS séparées.

- PLS(X,X)=ACP(X).

  Quand Y=X, les composantes t_k et u_k sont identiques. Le problème qui était de maximiser la covariance entre t et u, revient alors à maximiser la variance de t sous la contrainte $\|w\|^2$=1. Alors, $\lambda_1 = var(t_1)$ est la plus grande valeur propre de $V = X'DX$, matrice des covariances de X. On peut montrer que $\lambda_2 = var(t_2), ..., \lambda_k = var(t_k)$, pour tout $k = 1, ..., A$, et que, par récurrence, les w_k, qui sont de norme 1, sont les vecteurs propres associés aux valeurs propres λ_k de V classées en ordre décroissant. On retrouve donc l'Analyse en Composantes Principales de X.

Références

BAUMGARTNER, M. (1994), *Une introduction à Splus*, Ecole Polytechnique Fédérale de Lausanne.

BERTIER, P., et BOUROCHE, J.M. (1975), *Analyse des données multivariées*, Presses Universitaires de France.

BRY, X. (1996), *Analyses factorielles multiples*, Paris : Economica.

DURAND, J.F. (1997), *Additive Modeling of multivariate data by spline fonctions*, Rapport de Recherche n.97-04, Unité de Biométrie.

DURAND, J.F., SABATIER, R. (1997), Additive Splines for PLS regression, *Journal of the American Statistical Association*, Vol. 92, 440, 1546-1554.

MATHSOFT (1996), *S-Plus Version 3.4 for Unix Supplement*, Data Analysis Products Division, MathSoft, Seattle.

SWAYNE, D.F., COOK, D. & BUJA, A. (1991) *User's manual for XGobi, a Dynamic Graphics Program for Data Analysis Implemented in the X Window System (Release 2)*, Bellcore Technical Memorandum.

TENENHAUS, M. (1998) *La régression PLS, théorie et pratique*, Paris : Technip.

TENENHAUS, M. GAUCHI, J.P., et MÉNARDO, C. (1995), Régression PLS et applications, *Revue de Statistique Appliquée*, XLIII, 7-63.

WOLD, H. (1966), Estimation of principal component and related models by iterative least squares, *Multivariate Analysis*, ed.P.R.Krishnaiah, Newyork : Academic Press, 391-420.

WOLD, S. et al (1983), The multivariate calibration problem in chemistry solved by PLS method, *Proc. Conf. Matrix Pencils.* Ruhe, A. and Kagstrom, B. (Eds), Lecture notes in mathematics, Heidelberg : Springer Verlag, 286-293.