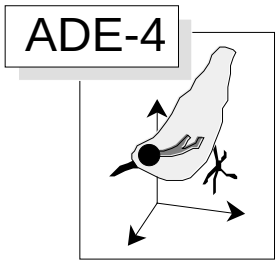


Régression PLS de seconde génération

Résumé

La fiche propose une introduction à la logique de la régression partiellement aux moindres carrés de deuxième génération. Quelques indications générales sont données sur des exemples reproductibles. Une comparaison des sorties du module PLS2gen avec celles du logiciel SIMCA (http://www.umetri.se/simca-pg.htm) utilisé dans un ouvrage à paraître de M. Tenenhaus est détaillée.

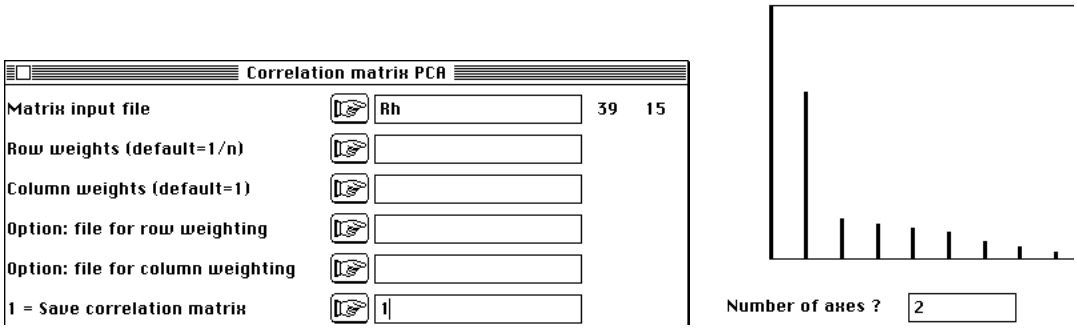
Plan

D. Chessel & L. Monimeau

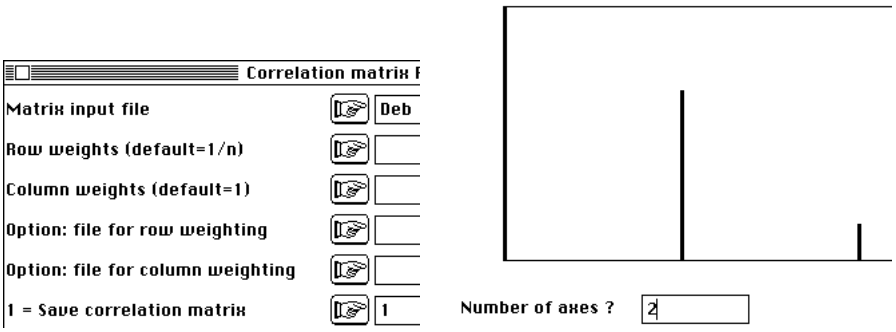
1 — Régressions simples et multiples

La régression PLS de seconde génération aborde la même situation que l'ACP sur variables instrumentales et on peut dire *grosso modo* que la PLS2 est à la régression PLS1 ce que l'ACPVI est à la régression multiple.

Utilisons les données de G. Carrel [1] sur les cartes Rhône et Rhône+1 de la pile ADE-4•Data. Le tableau Rh relève d'une ACP normée :

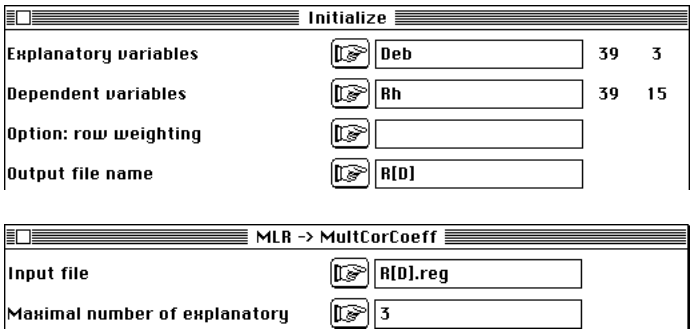


Le tableau Deb supporte le même programme :



Le premier est un tableau de variables à expliquer, le second est un tableau de variables explicatives. On peut évidemment envisager la régression multiple de chacune des 15 variables du premier groupe sur les 3 variables du second.

Examinons les carrés de corrélation (LinearReg : Initialize) puis (LinearReg : MLR -> MultCorCoeff) :



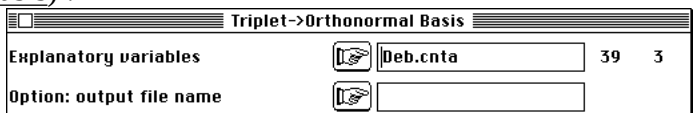
Le fichier R[D].yr2 est édité après transposition dans le tableur :

	1	2	3	1+2	1+3	2+3	1+2+3
1-Ta	**0.515**	0.064	0.059	0.519	0.533	0.341	0.584
2-Te	0.364	0.006	0.185	0.372	0.480	0.394	**0.529**
3-Co	**0.307**	0.044	0.072	0.311	0.341	0.317	0.419
4-pH	*0.021*	*0.018*	*0.021*	*0.031*	*0.050*	*0.024*	*0.051*
5-Ox	0.213	0.335	0.116	**0.436**	0.388	0.336	0.449
6-Tr	0.260	0.589	0.276	0.688	0.633	0.591	**0.729**
7-Dt	0.373	0.000	0.210	0.395	**0.509**	0.379	0.537
8-Dc	0.298	0.006	0.290	0.352	**0.509**	0.411	0.528
9-mg	0.400	0.248	0.023	**0.516**	0.463	0.296	0.516
A-Su	0.001	0.336	**0.565**	0.374	0.572	0.581	0.582
B-No	*0.113*	*0.083*	*0.015*	*0.155*	*0.118*	*0.245*	*0.255*
C-Ta	0.134	0.034	0.472	0.220	0.541	**0.583**	0.596
D-Ms	0.000	**0.736**	0.490	0.795	0.501	0.774	0.805
E-Mo	0.004	**0.625**	0.417	0.649	0.445	0.658	0.665
F-Ch	*0.034*	*0.222*	*0.219*	*0.225*	*0.286*	*0.269*	*0.299*

Le cas est intéressant car les régressions sont variées. La prévisibilité de la variable 4 est nulle (le pH qui forme le facteur 4 de l'ACP est quasiment une série aléatoire qui varie entre 6.8 et 7.2), certaines régression se font avec 1, 2 ou 3 variables explicatives. Les 3 explicatives jouent un rôle mais sont en plus corrélées. La question est : peut-on limiter le nombre de modèles à construire, les expliquées étant elles même fortement corrélées pour donner 2 facteurs d'ACP très nets (la variable 5 indépendante donne la composante 3 de l'ACP et n'est pas sans intérêt). L'ACPVI, d'un certain point de vue et la régression PLS2 dont c'est le but principal permettent de poursuivre cet objectif.

2 — Régression et variables instrumentales

Les variables explicatives définissent un sous-espace (Projectors : Triplet->Orthonormal Basis) :



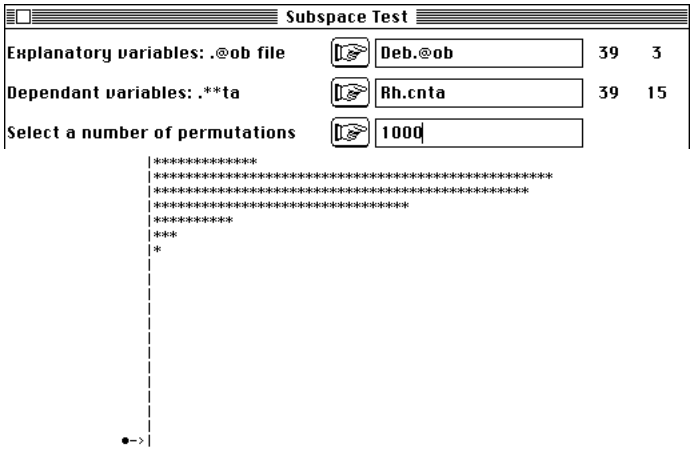

```
Orthonormalization: subspace generated by a statistical triplet
----------------------------------------
Explanatory variable file: Deb.cnta
It has 39 rows and 3 columns
----------------------------------------
Orthonormal basis: Deb.@ob
It has 39 rows and 3 columns
Row weight file: Deb.@pl
 (the same as Deb.cnpl)
Coordinates of the vectors of the orthonormal basis
in the basis of columns of Deb.cnta in : Deb.@co
File Deb.@co has 3 rows and 3 columns
----------------------------------------
```

La projections des variables dépendantes est exactement la régression multiple de chacune d'entre elles sur l'ensemble des explicatives et le pourcentage de variance expliquée n'est rien d'autre que le carré de corrélation :

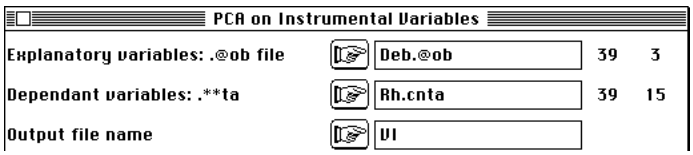
```
Projected inertia on a subspace
Orthonormal basis: Deb.@ob
It has 39 rows and 3 columns
Dependant variable file: Rh.cnta
It has 39 rows and 15 columns
```

	Subspace A	A Orthogo	Total	A+	A-
1	5.8370e-01	4.1630e-01	1.0000e+00	5836	4163
2	5.2920e-01	4.7080e-01	1.0000e+00	5292	4707
3	4.1895e-01	5.8105e-01	1.0000e+00	4189	5810
4	5.1146e-02	9.4885e-01	1.0000e+00	511	9488
5	4.4877e-01	5.5123e-01	1.0000e+00	4487	5512
6	7.2923e-01	2.7077e-01	1.0000e+00	7292	2707
7	5.3664e-01	4.6336e-01	1.0000e+00	5366	4633
8	5.2774e-01	4.7226e-01	1.0000e+00	5277	4722
9	5.1586e-01	4.8414e-01	1.0000e+00	5158	4841
10	5.8164e-01	4.1836e-01	1.0000e+00	5816	4183
11	2.5530e-01	7.4470e-01	1.0000e+00	2552	7447
12	5.9618e-01	4.0382e-01	1.0000e+00	5961	4038
13	8.0460e-01	1.9540e-01	1.0000e+00	8045	1954
14	6.6522e-01	3.3478e-01	1.0000e+00	6652	3347
15	2.9856e-01	7.0144e-01	1.0000e+00	2985	7014
Tot	7.5427e+00	7.4573e+00	1.5000e+01	5028	4971

Le test de la pertinence de cette projection ne s'impose pas (Projectors : Subspace Test) :



L'ACP sur variables instrumentales (Projectors : PCA on Instrumental Variables) est aisée :



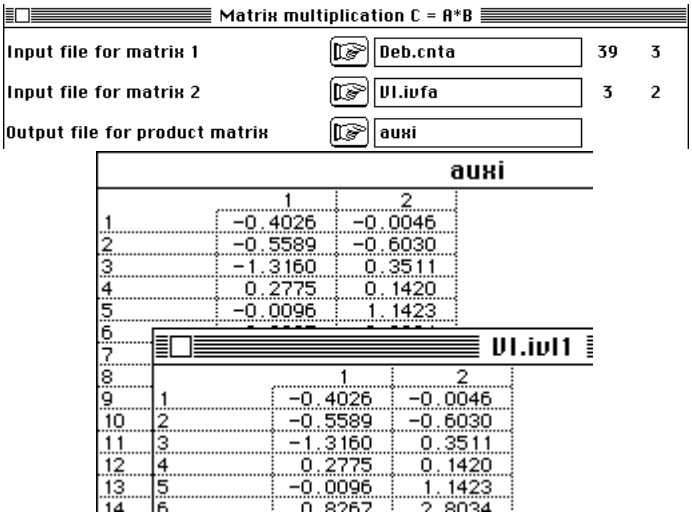
On garde 2 facteurs. On a deux systèmes d'interprétation.

```
| files VI.ivfa
|       VI.ivl1
|       VI.ivco
| allow a convenient interpretation
```

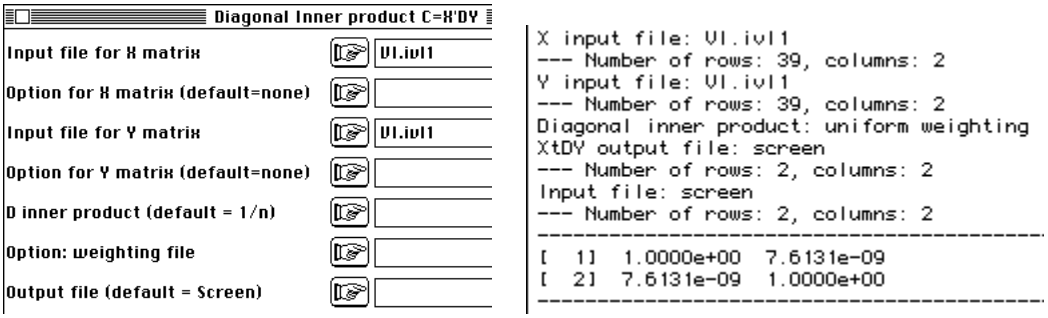
Les facteurs sont des poids pour les variables explicatives :

```
  1  0.61784    -0.17573
  2  0.71504    0.24392
  3  -0.36866   0.79972
```

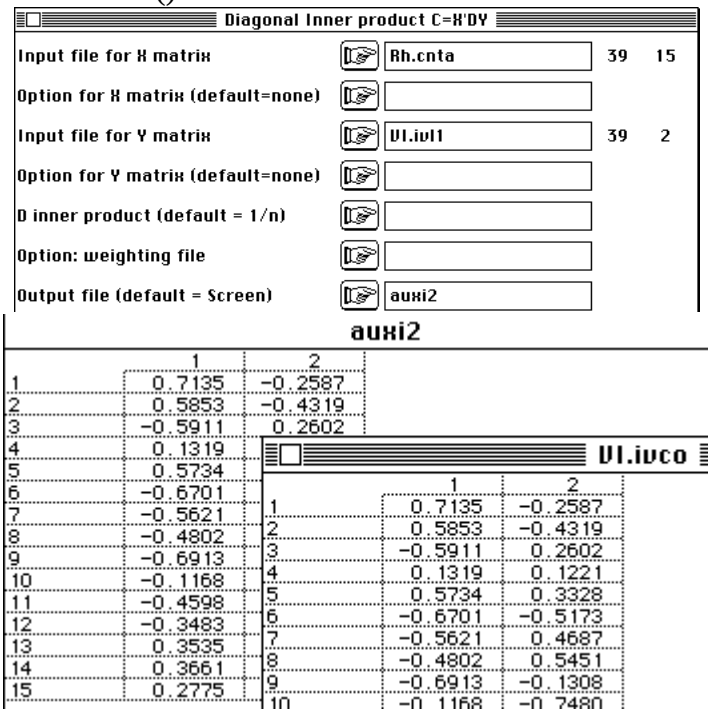
Ces poids sont utilisés pour calculer des combinaisons linéaires des explicatives. On peut vérifier par (MatAlg : Matrix multiplication C = A*B) :



Ces variables de synthèse sont de variance unité et de covariances nulles (MatAlg : Diagonal Inner product C=X'DY) :



Elles maximisent les sommes de carrés de corrélation avec les dépendantes. Ces corrélation sont dans VI.ivco () :



Les colonnes de VI.ivl1 sont donc des régresseurs communs de toutes les explicatives. En équation normalisée cela s'écrit que la variable :

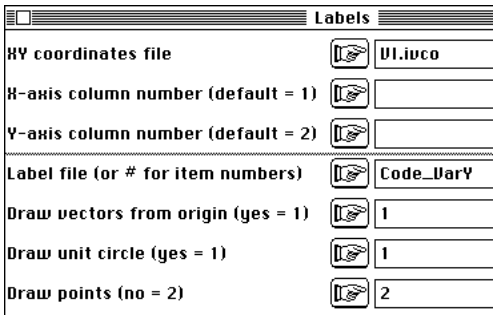
$$z^1 = 0.618\ x^1 + 0.715\ x^2 - 0.369\ x^3$$

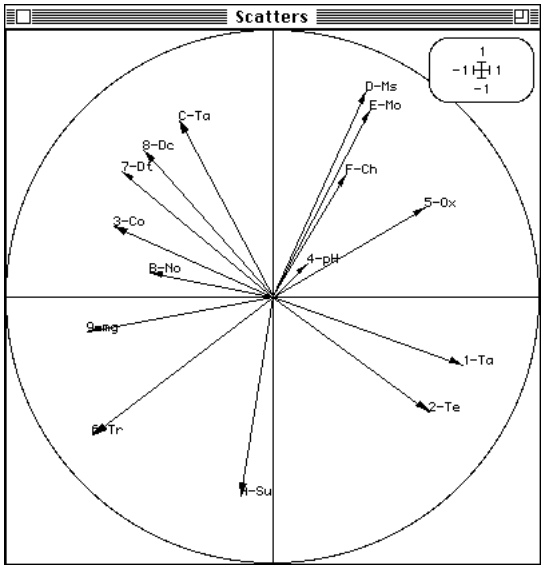
est un prédicteur simultané de toutes les dépendantes. On a un second prédicteur non corrélé au précédant avec :

$$z^2 = -0.176\ x^1 + 0.244\ x^2 - 0.800\ x^3$$

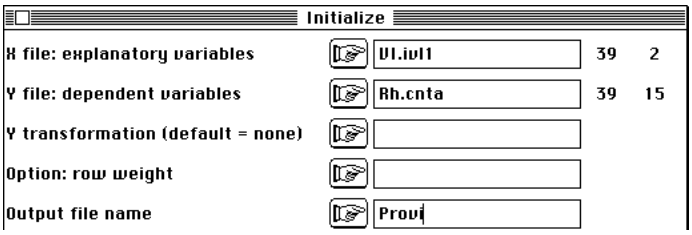
	1+2+3	z1	z2	z1+z2
1-Ta	0.584	**0.509**	0.067	0.576
2-Te	**0.529**	0.343	0.187	0.529
3-Co	0.419	0.349	0.068	0.417
4-pH	*0.051*	0.017	0.015	*0.032*
5-Ox	0.449	0.329	0.111	0.440
6-Tr	**0.729**	0.449	0.268	0.717
7-Dt	0.537	0.316	0.220	0.536
8-Dc	0.528	0.231	0.297	0.528
9-mg	0.516	**0.478**	0.017	0.495
A-Su	0.582	0.014	**0.560**	0.573
B-No	*0.255*	0.211	0.008	0.219
C-Ta	0.596	0.121	**0.433**	0.555
D-Ms	0.805	0.125	**0.592**	0.717
E-Mo	0.665	0.134	**0.488**	0.622
F-Ch	*0.299*	0.077	0.209	0.286

On peut élever les corrélations au carré (pourcentage de variance expliquée) puis les sommer (les explicatives de synthèse z^1 et z^2 sont indépendantes) pour obtenir les pourcentages expliquées par une régression multiple sur les mêmes explicatives. On est très proche de l'optimum. Le tout est dans la figure (Scatters : Labels) :



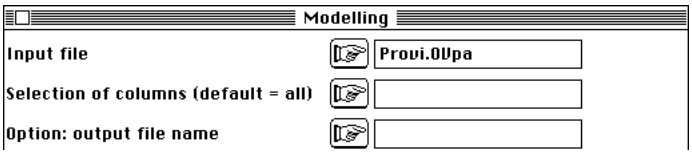


On retrouve les pourcentages de variance expliquée dans la régression orthogonale (OrthoVar : Initialize) :



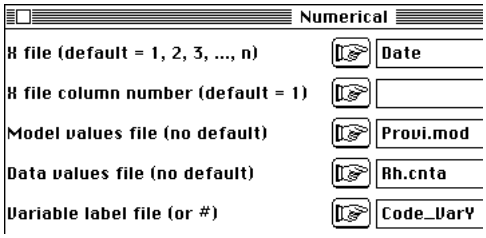

```
-------------------------------------------------------
New TEXT file Provi.OVpa contains the parameters:
----> Explanatory variables: VI.ivl1 [39][2]
----> Dependant variable file: Rh.cnta [39][15]
----> Transformation used: 0
      0 = None 1 = D-centring, 2 = D-standardization, 3 = D-normalization
----> Row weight file: Uniform_weighting
-------------------------------------------------------
File Provi.OVcs contains cosinus squared between explanatory and dependant
variables:
----> 2 rows (explanatory variables)
----> 15 columns (dependant variables
*-----------------------------------*
| N° | Variance | Explained|  Ratio   |
|   1| 1.000e+00| 5.760e-01| 5.760e-01|
|   2| 1.000e+00| 5.291e-01| 5.291e-01|
|   3| 1.000e+00| 4.171e-01| 4.171e-01|
|   4| 1.000e+00| 3.231e-02| 3.231e-02|
|   5| 1.000e+00| 4.396e-01| 4.396e-01|
|   6| 1.000e+00| 7.167e-01| 7.167e-01|
|   7| 1.000e+00| 5.357e-01| 5.357e-01|
|   8| 1.000e+00| 5.277e-01| 5.277e-01|
|   9| 1.000e+00| 4.951e-01| 4.951e-01|
|  10| 1.000e+00| 5.732e-01| 5.732e-01|
|  11| 1.000e+00| 2.191e-01| 2.191e-01|
|  12| 1.000e+00| 5.548e-01| 5.548e-01|
|  13| 1.000e+00| 7.172e-01| 7.172e-01|
|  14| 1.000e+00| 6.218e-01| 6.218e-01|
|  15| 1.000e+00| 2.860e-01| 2.860e-01|
*-----------------------------------*
```

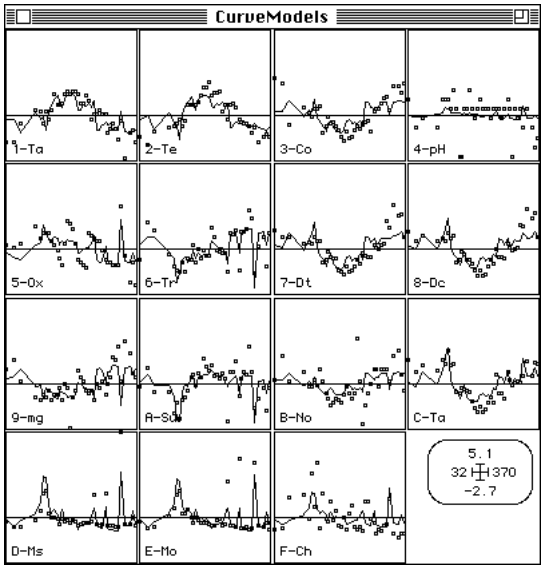
Modèles et résidus sont obtenus par OrthoVar: Modelling :



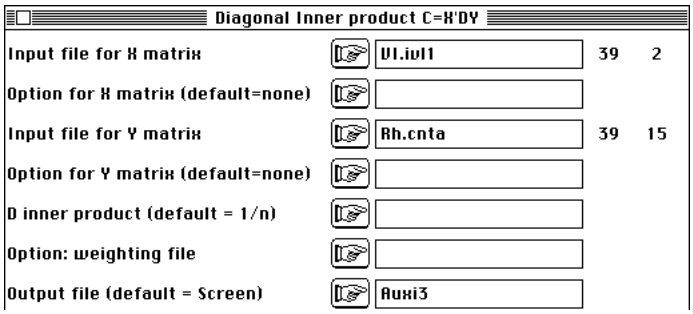

```
----> Explanatory variables: VI.ivl1
----> Dependant variable file: Rh.cnta
----> Transformation used: 0
      0 = None 1 = D-centring, 2 = D-standardization, 3 = D-normalization
----> Row weight file: Uniform_weighting
----> Selection of explanatory variables: 1a2
-------------------------------------------------------
File Provi.mod has 39 rows and 15 columns
It contains linear models
from separate multiple linear regression of each dependant variable
upon the set of explanatory variables
File :Provi.mod
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.042e+00| 1.204e+00|
|   2|-1.434e+00| 1.187e+00|
•••
|  14|-8.134e-01| 2.490e+00|
|  15|-5.871e-01| 1.679e+00|
|----|----------|----------|
```

Données et modèles sont représentées par CurveModels : Numerical :

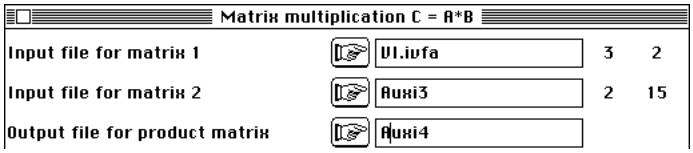




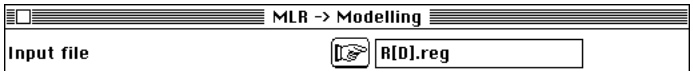
Comme les régresseurs sont de variances unité et de covariances nulles, les coefficients de régression sont de simples coefficients de corrélation, donc de simples produits scalaires (MatAlg : Diagonal Inner product C=X'DY) :



Le fichier Auxi3 contient les coefficients de régression des variables y (en colonnes) sur les deux variables z en lignes. Le produit de matrice entre les fichiers VI.ivfa (3-2) et Auxi3 (2-15) donne les équations de régression dans les variables de départ (variables normalisées). Utiliser MatAlg : Matrix multiplication C = A*B :

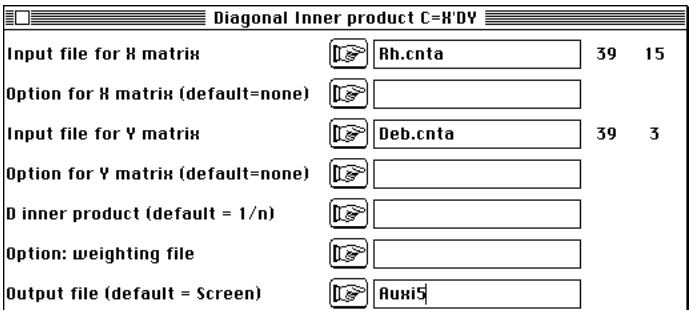


On édite le résultat (après transposition) et on le compare aux coefficients obtenus par régression multiple directe et aux corrélations explicatives-dépendantes. On a les premiers par LinearReg : MLR -> Modelling :



```
File R[D].MLRw1 has 3 rows and 15 columns
It contains regression coefficients
Rows : explanatory variables / Columns : dependant variables
Models for normalized (mean = 0 / variance =1) variables
File :R[D].MLRw1
|Col.|   Mini    |   Maxi    |
|----|-----------|-----------|
|   1|-3.686e-01| 5.709e-01|
|   2|-5.751e-01| 4.258e-01|
|   3|-4.150e-01| 4.762e-01|
|   4|-5.202e-02| 2.074e-01|
|   5| 1.655e-01| 3.882e-01|
```

Les seconds sont des produits scalaires (MatAlg : Diagonal Inner product C=X'DY) :



On regroupe les résultats :

	Genève	Arve	Autres	Genève	Arve	Autres	Genève	Arve	Autres
1-Ta	0.486	0.447	-0.470	0.571	0.334	-0.369	0.718	0.252	-0.243
2-Te	0.438	0.313	-0.561	0.426	0.329	-0.575	0.603	0.075	-0.430
3-Co	-0.411	-0.359	0.426	-0.369	-0.415	0.476	-0.554	-0.210	0.267
4-pH	0.060	0.124	0.049	0.192	-0.052	0.207	0.146	0.133	0.144
5-Ox	0.296	0.491	0.055	0.388	0.368	0.165	0.462	0.579	0.341
6-Tr	-0.323	-0.605	-0.167	-0.431	-0.462	-0.296	-0.510	-0.767	-0.525
7-Dt	-0.430	-0.288	0.582	-0.460	-0.247	0.546	-0.611	-0.022	0.459
8-Dc	-0.393	-0.210	0.613	-0.396	-0.205	0.608	-0.546	0.077	0.538
9-mg	-0.404	-0.526	0.150	-0.543	-0.341	-0.016	-0.633	-0.498	-0.151
A-Su	0.059	-0.266	-0.555	-0.029	-0.148	-0.661	0.033	-0.579	-0.751
B-No	-0.299	-0.307	0.240	-0.116	-0.552	0.459	-0.336	-0.289	0.124
C-Ta	-0.331	-0.088	0.655	-0.135	-0.350	0.890	-0.367	0.184	0.687
D-Ms	0.083	0.440	0.485	-0.202	0.820	0.144	-0.003	0.858	0.700
E-Mo	0.103	0.432	0.424	-0.097	0.699	0.183	0.063	0.791	0.646
F-Ch	0.091	0.310	0.263	0.199	0.166	0.392	0.183	0.471	0.468
	ACPVI			**MLR**			**Corrélations**		

Tableau 1 : Liaisons entre les 15 variables dépendantes et les 3 variables explicatives vues de trois manières différentes. A gauche, coefficients de régression après une ACPVI, au centre coefficients de régression après une régression multiple et à droite coefficients de corrélation ordinaire.

Les trois points de vue sont cohérents. On peut résumer les opérations par la figure 1. Nous sommes typiquement devant un cas de régression multiple (MLR) ne présentant aucune pathologie propre à cette méthode. Bien que corrélées, les explicatives font entre elles des angles suffisants pour que les sous-espaces de projection soient stables. Il n'y a pas d'incohérence entre corrélations et coefficients de régression. Il y a peu d'écarts entre projections sur les variables et projections sur les composantes, c'est-à-dire entre projection sur le sous-espace des explicatives et sur une partie de celui-ci choisie d'une manière ou d'une autre. C'est loin d'être le cas le plus fréquent mais le jeu de données permet de poser la question : la régression PLS introduite ici permet-elle de savoir si on peut s'en passer, quand la régression multiple ordinaire est acceptable ?

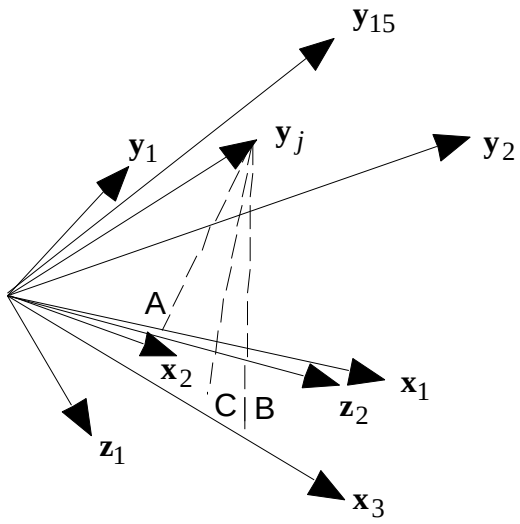


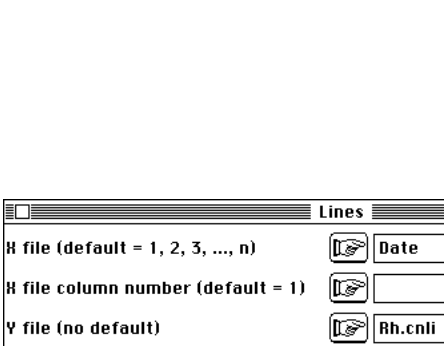
*Figure 1 : Régression et projections sur des variables normalisées. Les vecteurs **x** sont des variables explicatives et les vecteurs **y** sont des variables dépendantes.*
Premier point de vue — A : une régression simple est la projection d'un vecteur sur un autre. Coefficient de régression ou coefficient de corrélation sont confondus.
Second point de vue — Les explicatives forment un sous-espace vectoriel. B : une régression multiple est la projection d'une explicative sur ce sous-espace. La projection est toujours possible, le cosinus carré de l'angle entre le vecteur et son projeté est le carré de corrélation. Comme combinaison des explicatives le projeté donne les coefficients de régression.
*Troisième point de vue — Les vecteurs **z** sont une base orthonormée d'un sous-espace du précédent. Si on prend les composantes principales de l'ACP des explicatives, on retrouve la régression sur composantes. Si on prend les composantes principales du nuage des vecteurs projetés on retrouve la régression par ACPVI. C : Projection de la dépendante sur ce sous-espace.*

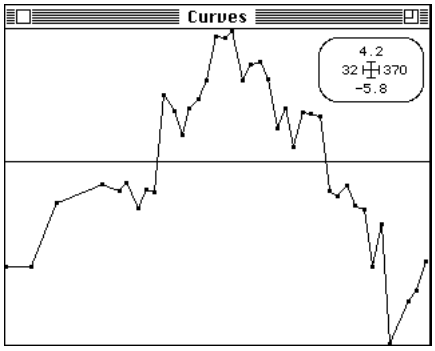
3 — Régressions sur composantes

Nous venons de voir que le régression peut se faire par projection sur l'espace des variables explicatives ou par projection sur un sous-espace de celui-ci. Quel est l'intérêt de cette complication ? Il y a deux objectifs sous-jacents. Le premier est la recherche de modèles communs à toutes les dépendantes, le second est la recherche de modèles numériquement plus stables.

3.1 — Auto-modélisation

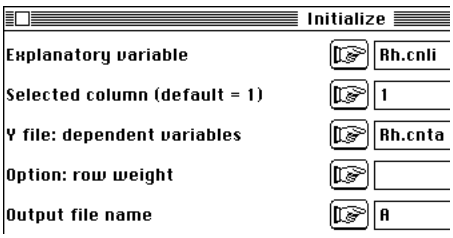
On peut considérer que plusieurs variables peuvent être des images d'un même phénomène et qu'elles peuvent supporter une prévision par un même régresseur. C'est assez clair dans le cas de l'auto-modélisation des données par elle-même. Considérons par exemple la première coordonnée de l'ACP des variables dépendantes. Cette courbe définit un effet saisonnier :

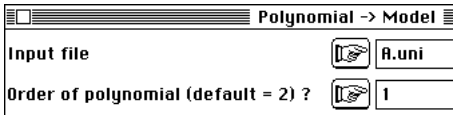




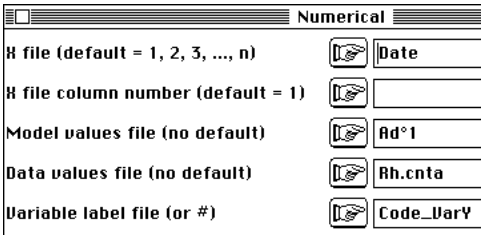
Il est logique de chercher à prédire les données par cette courbe unique : c'est le principe de reconstitution des données. Sachant que la première coordonnée est la variable de synthèse (définie à une constante multiplicative près) qui optimise la somme des corrélations avec les variables de départ [2], on ne peut espérer de régresseur commun plus efficace.

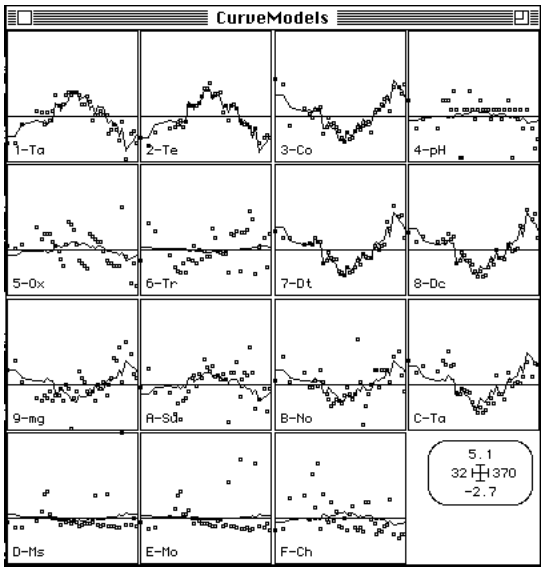
On peut faire cette régression directement (UniVarReg : Initialize et UniVarReg : Polynomial -> Model) :





ce qui donne :

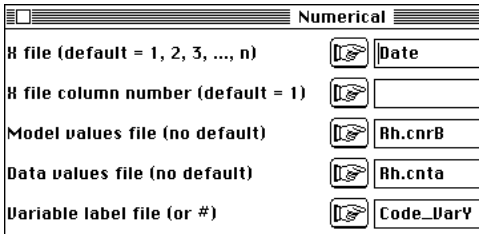


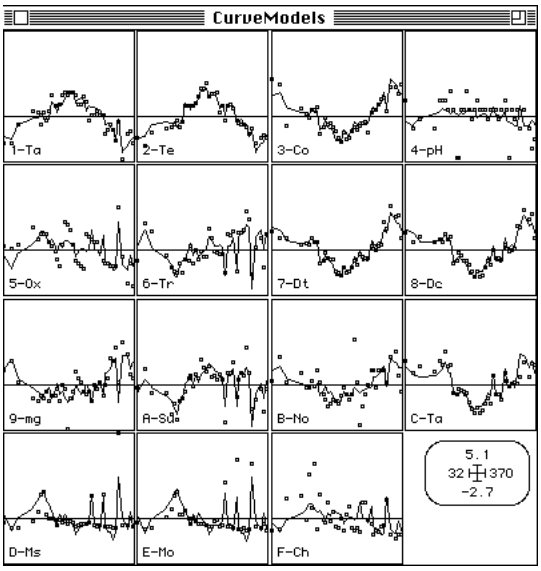


On obtient le même résultat avec DDUtil : Data modelling. On a bien ici un modèle commun, à savoir la même courbe de référence ajustée à plusieurs variables à l'aide d'un seul coefficient multiplicatif. Ce coefficient peut être positif ou négatif et éventuellement nul lorsque la variable n'a rien à voir avec le régresseur commun.

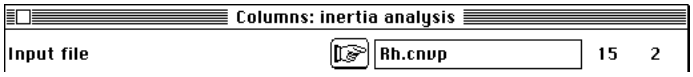
Par contre, quand on prend deux régresseurs communs comme les deux premières coordonnées de l'ACP, bien que les variables qui génèrent les modèles soient communes, le simple fait de les utiliser de manière variable génèrent des modèles qui peuvent être très variés. On peut faire cela directement ou utiliser DDUtil : Data modelling :


```
Modelling: Data reconstitution after PCA/CA
Title of the analysis: Rh.cnta
Number of rows: 39, columns: 15

File Rh.cnrA contains the model computed with 1 axe
It has 39 rows and 15 columns

File Rh.cnrB contains the model computed with 2 axes
It has 39 rows and 15 columns
```





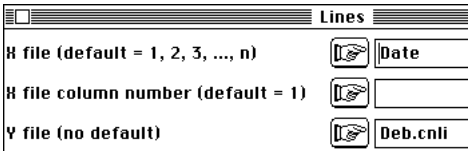
On reconstitue ainsi une grande part de la variabilité des données (DDUtil : Columns/Inertia analysis):

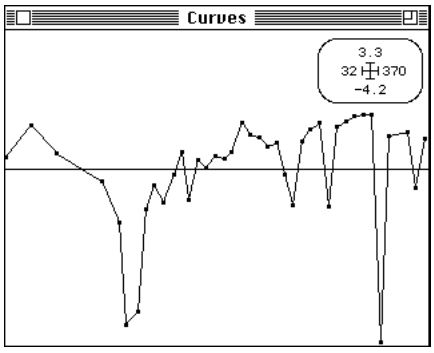



```
----------Relative contributions----------
|Num  |Fac 1|Fac 2||Remains| Weight | Cont.|
|    1| 8201|  344|| 1454  |10000   |  666 |
|    2| 9311|   34||  654  |10000   |  666 |
|    3| 8690|  130|| 1178  |10000   |  666 |
|    4|  372|  703|| 8923  |10000   |  666 |
|    5|  769| 4227|| 5002  |10000   |  666 |
|    6|  129| 7528|| 2342  |10000   |  666 |
|    7| 9355|   88||  555  |10000   |  666 |
|    8| 8894|  474||  630  |10000   |  666 |
•••
```

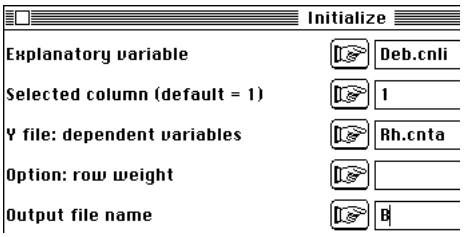
3.2 — Régression sur composantes principales

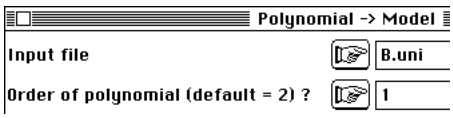
Cette logique des régresseurs communs s'applique aux coordonnées d'une ACP d'un tableau extérieur (régression sur composantes ou PCR). La même idée est en jeu. En prenant une composante on a un modèle commun et en en prenant deux on a des modèles stables puisqu'on élimine ainsi des variables explicatives une part de leur variabilité aléatoire. La première coordonnée du tableau des explicatives a la forme :



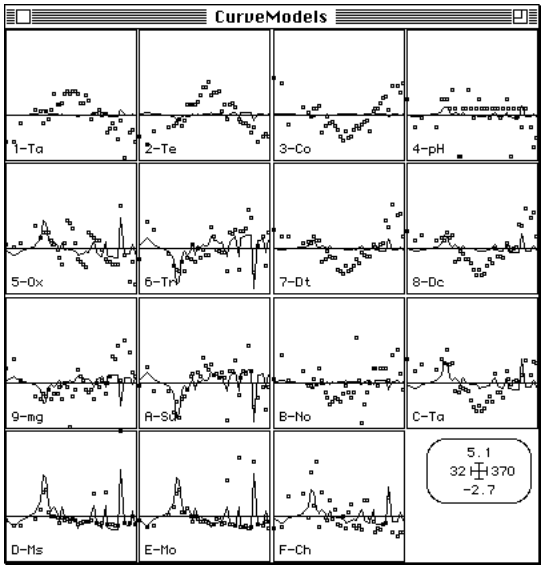


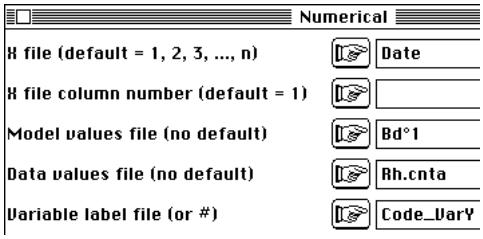
On peut faire cette régression directement (UniVarReg : Initialize et UniVarReg : Polynomial -> Model) :



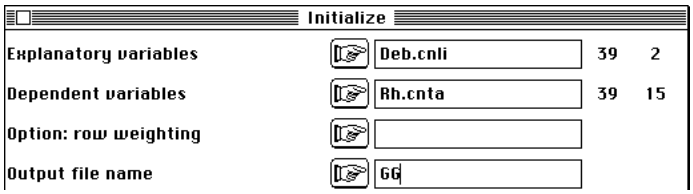


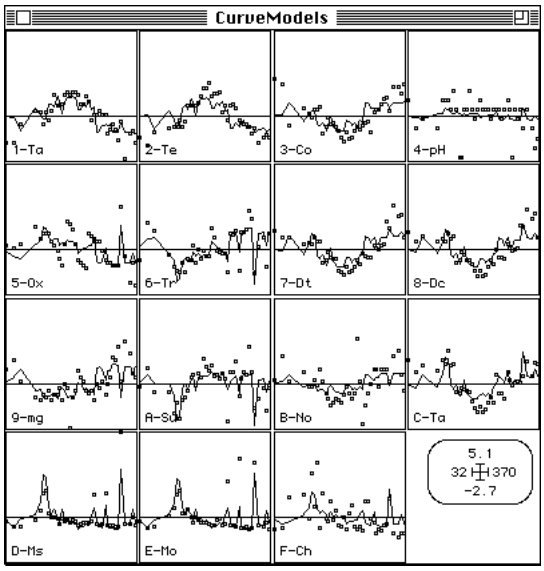
ce qui donne :

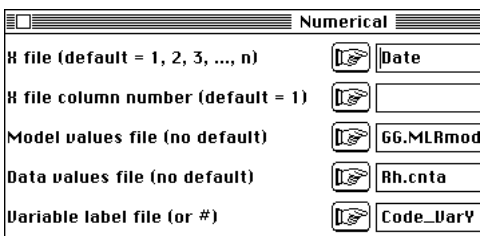




On note l'inversion de priorité : la première composante externe (rôle du débit des affluents sur les apports de matériel minéral et organique) joue le rôle de la seconde composante interne. On refait la même observation sur la diversité des modèles construits sur deux régresseurs communs (LinearReg : Initialize) :



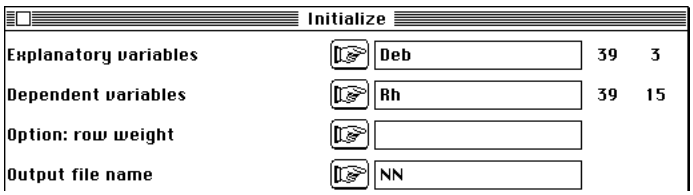




On pourra chercher encore à comparer les résultats avec les précédentes régression tant du point de vue des coefficients de régression que du point de vue de la qualité des résultats. Toutes ces approches sont ici destinées à introduire l'idée de régressions sur composantes qui font précéder la régression proprement dite par une recherche des prédicteurs. C'est l'idée fondamentale des régressions PLS qui soulève la question première du nombre de composantes à utiliser.

3.3 — Le nombre de composantes PLS

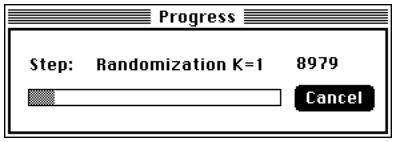
La régression PLS est une méthode de prédiction linéaire qui utilise la notion de composantes explicatives et permet de discuter de la construction progressive du modèle par adjonction progressives de composantes. La question du nombre de composantes à introduire est donc fondamental. Le module d'ADE-4 propose un test de permutations très simple. Initialiser par PLSgen2 : Initialize :



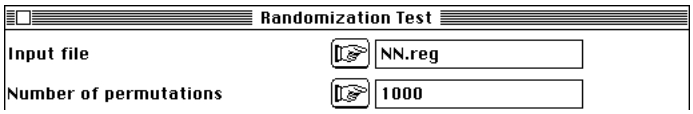

```
-------------------------------------------------------
New TEXT file NN.reg contains the parameters:
----> Explanatory variables: Deb [39][3]
----> Dependant variable file: Rh [39][15]
----> Row weight file: Uniform weight
-------------------------------------------------------
```

Lancer le test par PLSgen2 : Randomization Test :

Si on juge avoir présumé de sa patience :

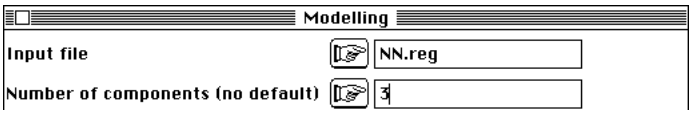



```
Error: Computations cancelled by user
```



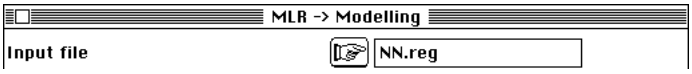

```
Explanory variable file: Deb
It has 39 rows and 3 columns
-----------------------------------------
Dependent variable file: Rh
It has 39 rows and 15 columns
-------------------------------------
-------- Dependent variable Y ---------
-------------------------------------
| Step | Nrepet | X>Xobs | Frequencie |
|     1|    1000|       0|   0.000e+00|
|     2|    1000|       0|   0.000e+00|
|     3|    1000|      41|   4.100e-02|
```

A chaque pas, l'introduction d'une nouvelle composante est testée par un ensemble de permutations sur le tableau des explicatives prenant en compte les composantes précédentes. Ici, les deux premières sont fortement significatives et la troisième peut encore être considérée comme acceptable. Retenir trois composantes pour trois explicatives, c'est affirmer que la régression multiple est valide. Dans le modèle, on trouve donc exactement le même résultat par les deux méthodes. PLSgen2 : Modelling donne :




```
-----------------------------------------------
File NN.PLSgen2mod contains components model
It has 39 rows and 15 columns
File :NN.PLSgen2mod
|Col.|    Mini    |    Maxi    |
|----|------------|------------|
|   1| 6.140e+00| 2.250e+01|
|   2| 5.019e+00| 1.963e+01|
|   3| 2.440e+02| 3.187e+02|
|   4| 7.980e+00| 8.174e+00|
|   5| 8.486e+01| 1.017e+02|
|   6|-1.910e+01| 1.843e+02|
|   7| 1.258e+02| 1.961e+02|
|   8| 4.199e+01| 7.001e+01|
|   9| 5.008e+00| 7.482e+00|
|  10| 1.928e+01| 4.077e+01|
|  11| 3.688e-01| 6.715e-01|
|  12| 1.115e+02| 2.126e+02|
|  13|-7.859e+00| 1.879e+02|
|  14| 6.267e-01| 1.580e+01|
|  15| 1.576e+00| 7.106e+00|
|----|------------|------------|
-----------------------------------------------
```

LinearReg : MLR -> Modelling propose :



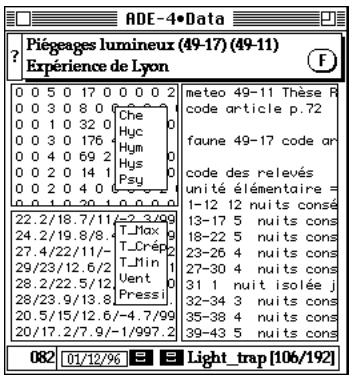

```
-----------------------------------------
File NN.MLRmod has 39 rows and 15 columns
It contains linear models
from separate multiple linear regression of each dependant variable
upon the set of explanatory variables
File :NN.MLRmod
|Col.|    Mini    |    Maxi    |
|----|------------|------------|
|   1| 6.140e+00| 2.250e+01|
|   2| 5.019e+00| 1.963e+01|
|   3| 2.440e+02| 3.187e+02|
|   4| 7.980e+00| 8.174e+00|
|   5| 8.486e+01| 1.017e+02|
|   6|-1.910e+01| 1.843e+02|
|   7| 1.258e+02| 1.961e+02|
|   8| 4.199e+01| 7.001e+01|
|   9| 5.008e+00| 7.482e+00|
|  10| 1.928e+01| 4.077e+01|
|  11| 3.688e-01| 6.715e-01|
|  12| 1.115e+02| 2.126e+02|
|  13|-7.859e+00| 1.879e+02|
|  14| 6.267e-01| 1.580e+01|
|  15| 1.576e+00| 7.106e+00|
|----|------------|------------|
-----------------------------------------
```

La régression PLS est donc une régression linéaire qui valide la méthode classique quand celle-ci est légitime.

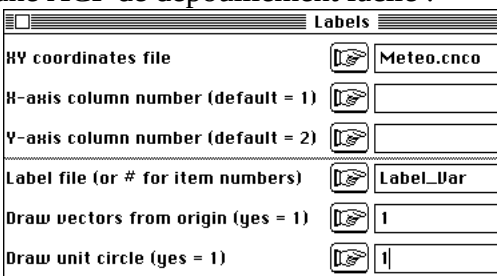
Dans les cas contraires elle apporte des éléments décisifs.

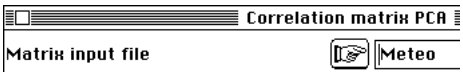
4 — La première composante PLS

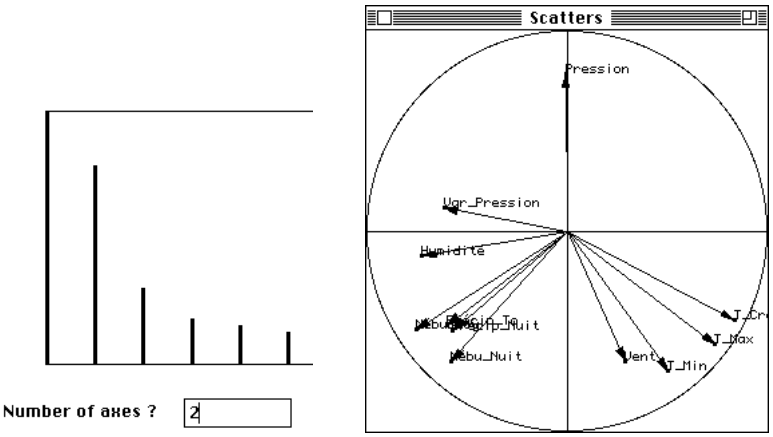
On utilise les données crées par la carte Light_trap [3] de la pile ADE-4•Data :

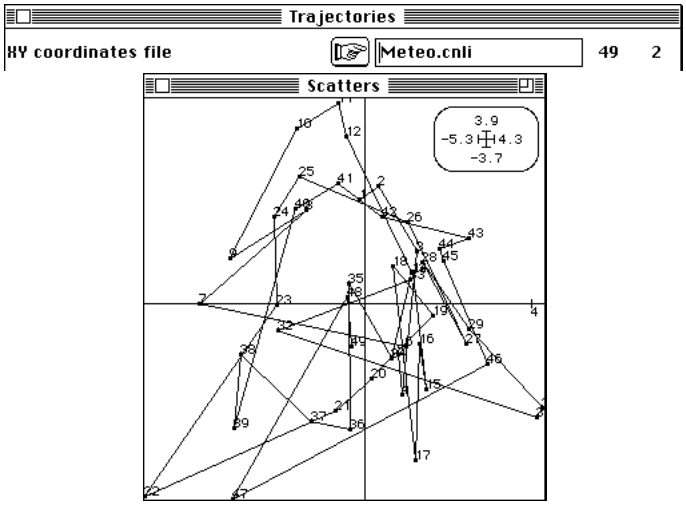


Les variables explicatives (11) donnent une ACP de dépouillement facile :

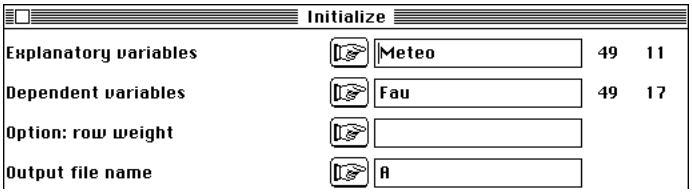




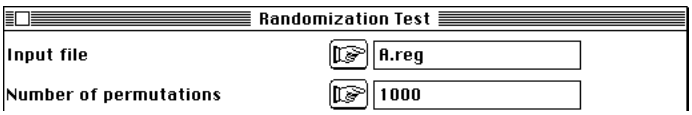




On note la succession haute pression (beau temps) puis fortes températures puis précipitations (orages d'été) caractéristiques du temps estival de la région. La question porte sur l'influence des variables météorologique sur l'abondance des piégeages lumineux. le tableau faunistique a 17 espèces (variables). Le nombre élevé d'explicatives (11) pour 49 relevés et le nombre élevé de modèles à construire (17) invite fortement à une régression sur composantes. On utilise la PLS :




```
New TEXT file A.reg contains the parameters:
----> Explanatory variables: Meteo [49][11]
----> Dependant variable file: Fau [49][17]
----> Row weight file: Uniform weight
```




```
Explanory variable file: Meteo
It has 49 rows and 11 columns
-----------------------------------------
Dependent variable file: Fau
It has 49 rows and 17 columns
-----------------------------------------
-------- Dependent variable Y ---------
-----------------------------------------
| Step | Nrepet | X>Xobs | Frequencie |
|    1|    1000|       2|   2.000e-03|
|    2|    1000|     526|   5.260e-01|
```

La première composante est très significative, la seconde ne l'est pas du tout et le test s'arrête automatiquement. Il existe un régresseur commun.

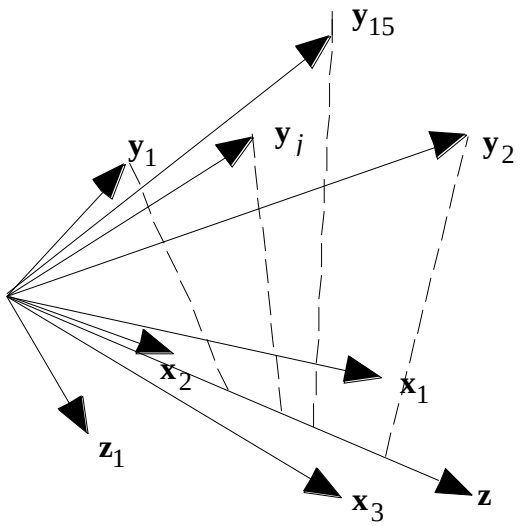
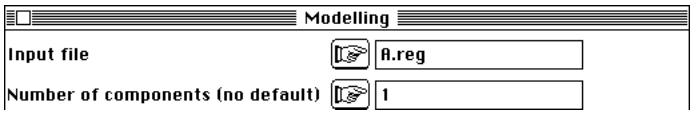


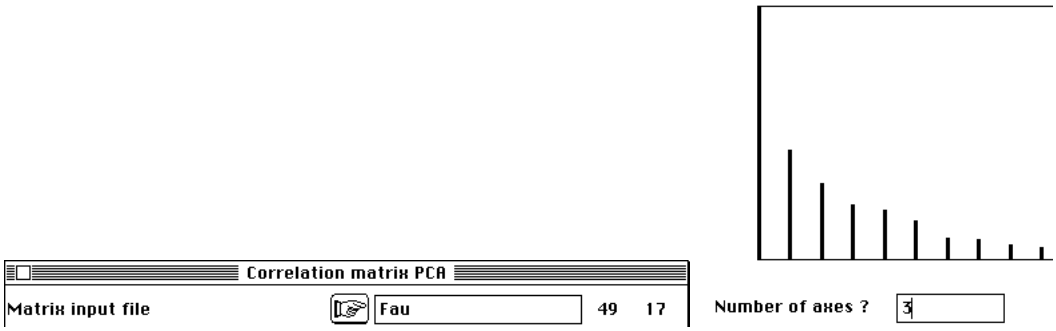
*Figure 2 : Régression PLS à une composante explicative. Les vecteurs **x** sont des variables explicatives et les vecteurs **y** sont des variables dépendantes. **z** est une composante explicative qui sert de prédicteur unique.*



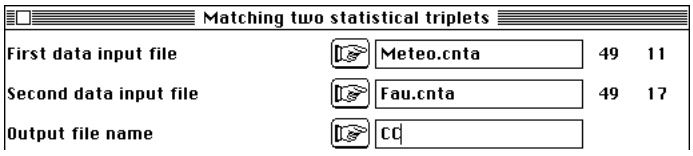
Le programme édite toutes les corrélations entre toutes les variables en jeu (explicatives et dépendantes) :

```
---------------------- Correlation matrix ------------------
[  1] 1000
[  2]  882 1000
[  3]  766  770 1000
[  4]  521  478  378 1000
[  5] -371 -246 -488 -449 1000
[  6] -448 -500 -271 -199   97 1000
[  7] -410 -512 -213 -125  -55  265 1000
[  8]  -94 -274  123  284 -568  211  410 1000
[  9]  -96 -147   72   36 -188  226  399  444 1000
•••
```

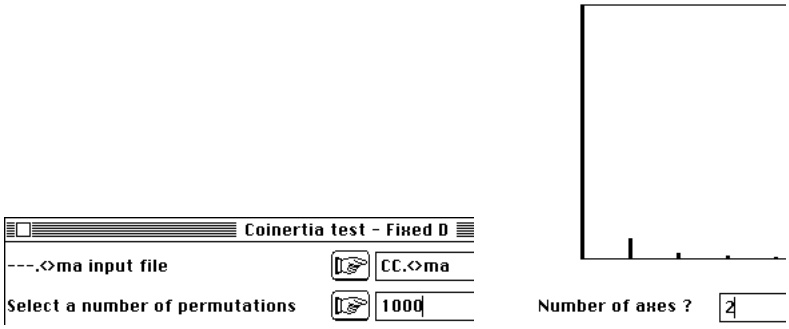
puis il fournit de nombreuses indications. Le principe de fonctionnement à l'aide d'une seule composante est relativement simple. La méthode est fortement liée à l'analyse de co-inertie et on peut ici explorer les résultats. On a déjà fait l'ACP normée du tableau Meteo. On peut faire de même pour le tableau Fau.

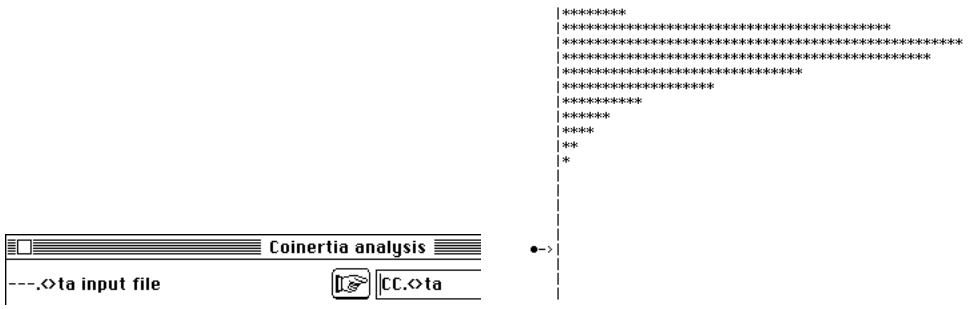


On peut alors exécuter l'analyse de co-inertie :



Vérifier l'existence d'une co-structure et exécuter l'analyse :





On a trouvé une combinaison des explicatives et une combinaison des expliquées de covariance maximale et le premier facteur de co-inertie est capital :

Num	Covaria.	Varian1	varian2	Correla.	INER1	INER2
1	3.142	3.659	6.024	0.6691	4.049	6.278
2	0.8897	2.203	1.802	0.4465	3.188	2.738

La première coordonnée sur l'axe de co-inertie du tableau Meteo (CC.<>l1) se retrouve dans la première colonne de A.PLSgen2.t. Les 3 autres colonnes sont sans signification. On trouve donc dans A.PLSgen2.t les composantes explicatives :

N.B. La chaîne .PLSgen2 a été changée en _PLS2 dans les noms de fichiers

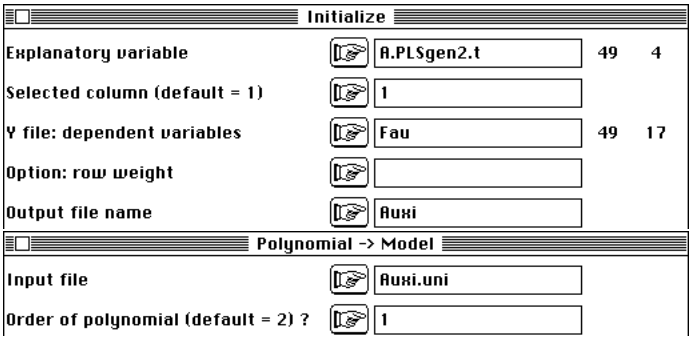
A.PLSgen2.t

	1
1	0.6818
2	0.7040
3	-0.4750
4	-1.2746
5	-1.2713
6	-1.0730
7	3.4907
8	2.1051
9	3.1936
10	2.7734
11	2.3923
12	1.9889
13	-0.4102
14	-0.7031
15	-2.3871
16	-1.4059
17	-2.5083
18	-0.1148
19	-1.1715
20	-0.5686
21	0.0766
22	2.7361
23	1.8721
24	2.2093
25	2.2982

CC.<>l1

	1	2
1	0.6818	-1.4526
2	0.7040	-1.3772
3	-0.4750	-0.7413
4	-1.2746	2.1872
5	-1.2713	0.9788
6	-1.0730	0.2933
7	3.4907	1.4907
8	2.1051	0.2266
9	3.1936	1.0201
10	2.7734	-1.3742
11	2.3923	-2.3660
12	1.9889	-2.4035
13	-0.4102	-0.5794
14	-0.7031	-0.9116
15	-2.3871	1.2667
16	-1.4059	0.0481
17	-2.5083	2.3961
18	-0.1148	-0.7859
19	-1.1715	-0.5733
20	-0.5686	1.6856
21	0.0766	1.8061
22	2.7361	1.7167
23	1.8721	1.0234

La modélisation du tableau Fau se fait donc par régression simple sur une combinaison de variables explicatives (UniVarReg) :



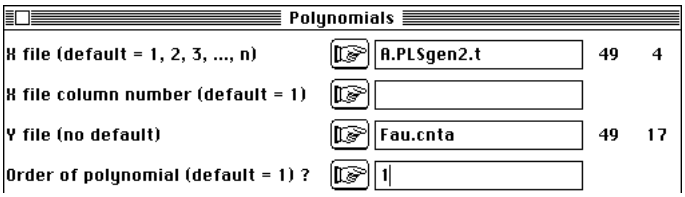
A.PLSgen2mod

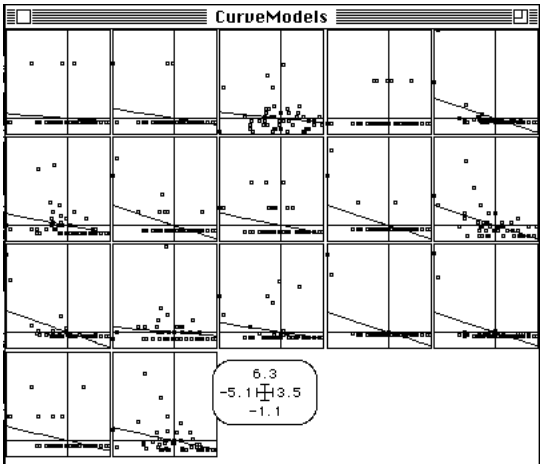
	1	2	3	4	5
1	0.0507	0.0397	3.5124	0.1228	42.4390
2	0.0504	0.0390	3.5052	0.1228	39.8443
3	0.0685	0.0763	3.8896	0.1222	177.7789
4	0.0809	0.1015	4.1503	0.1219	271.3189

Auxid°1

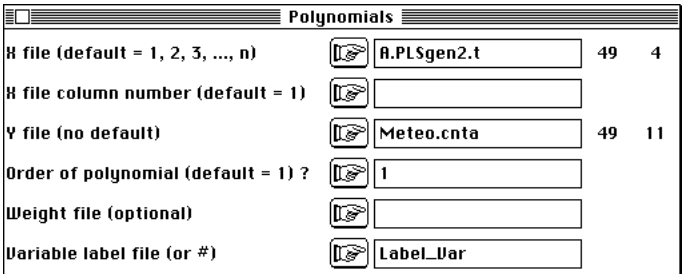
	1	2	3	4	5
1	0.0507	0.0397	3.5124	0.1228	42.4390
2	0.0504	0.0390	3.5052	0.1228	39.8443
3	0.0685	0.0763	3.8896	0.1222	177.7789
4	0.0809	0.1015	4.1503	0.1219	271.3190
5	0.0808	0.1014	4.1492	0.1219	270.9368

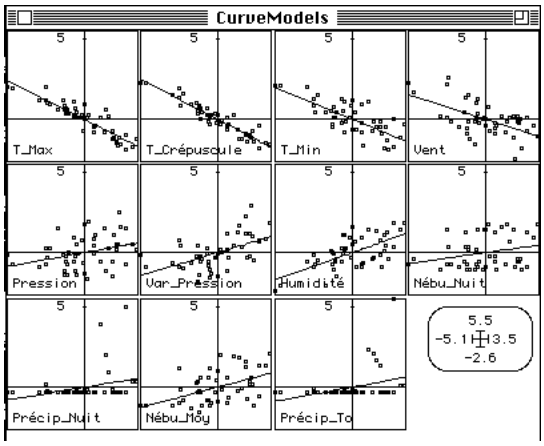
Pour juger de l'adéquation du modèle on préférera la reconstitution du tableau normalisé directement sur le graphique (CurveModels) :





On voit immédiatement que la composante explicative a une corrélation de signe constant avec toutes les expliquées et que de faibles valeurs de la composante explicative correspondent à une abondance accrue de toutes les espèces. La constitution de cette composante est explicitée par :





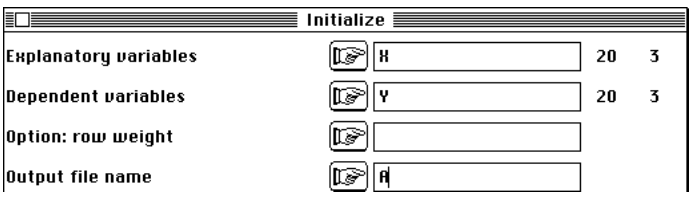
Les envols de toutes les espèces sont favorisées par des températures élevées, des pressions atmosphériques faibles et en baisse et une absence de précipitations, donc le moment qui précède l'éclatement des orages estivaux. Le graphe portant sur la faune montre également que la régression linéaire est fort mal adaptée aux dénombrements. On aurait amélioré la situation par une transformation en log mais une bonne partie de l'information est en présence-absence (valeurs à droite des graphiques) et la notion de niche (l'abondance est une distribution de fréquence) est nettement meilleure que la notion d'abondance normalisée (ce qui est bien connue).

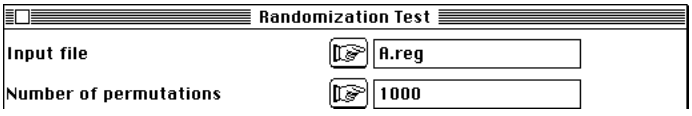
On retiendra qu'une PLS de seconde génération est, en cas d'une composante unique, une régression simple sur la première coordonnée de l'analyse de co-inertie (tableau des explicatives) qui est ici une analyse inter-batterie de Tucker [4].

5 — Composantes explicatives multiples

Pour tout savoir sur la PLS, le plus simple est d'utiliser l'ouvrage en préparation de M. Tenenhaus [5], ouvrage qui font suite aux documents disponibles [6]. La plus grande transparence méthodologique y est disponible et permet de situer le module d'ADE-4 par rapport au logiciel le plus utilisé de la chimiométrie à savoir SIMCA[7]. La grande différence porte sur les données manquantes. Elles ne son pas tolérées par PLSgen2. Le fonctionnement interne du programme en est très différent mais les résultats sont identiques dans le cas de données complètes.

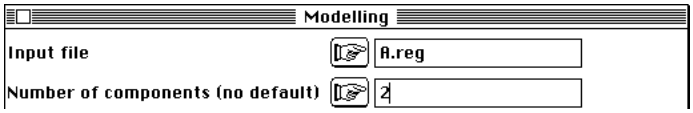
Pour comparer les résultats obtenus par M. Tenenhaus avec SIMCA-P et ceux fournis par PLSgen2 utiliser la carte Lineerud de la pile ADE-4•Data. Les renvois à l'ouvrage cité de M. Tennehaus se font par *ibidem* suivi du numéro de page.





Step	Nrepet	X>Xobs	Frequencie
1	1000	58	5.800e-02
2	1000	613	6.130e-01

La première composante est significative faiblement, la seconde ne l'est pas du tout et nous gardons deux composantes uniquement pour vérifier les calculs :



Ces calculs sont dépourvus de signification expérimentale. On retrouve la matrice des corrélations (*ibidem* p. 23) avec 1-2-3 variables explicatives, 4-5-6 variables dépendantes :

```
---------------------- Correlation matrix -------------------
[  1] 1000
[  2]  870 1000
[  3] -366 -353 1000
[  4] -390 -552  151 1000
[  5] -493 -646  225  696 1000
[  6] -226 -191   35  496  669 1000
-------------------------------------------------------------
```

Les composantes explicatives notée t_k sont des vecteurs, combinaisons linéaires de variables explicatives centrées et de variance unité. Par raison de cohérence globale les variances dans PLSgen2 sont calculées avec la pondération uniforme ($1/n$) et celles de SIMCA sont calculées dans le cadre gaussien ($1/n-1$).

A.PLSgen2.t

	1	2	3
1	-0.6596	0.5977	-0.1067
2	-0.7897	0.1682	0.1099
3	-0.9310	-0.5267	0.0388
4	0.7063	-0.6875	0.2831
5	-0.4994	1.1447	-0.1484
6	-0.2350	-0.0724	0.0202
7	-1.4402	-0.0768	-0.4689
8	0.7629	-0.2128	-0.0265
9	1.7596	-0.6598	-1.2772
10	1.1928	0.1681	0.2730
11	0.3740	0.7077	0.1653
12	0.7627	0.7055	0.0023
13	1.2175	-0.7652	0.2753
14	-4.5038	-0.7682	0.2090
15	-0.8446	0.9840	-0.0672
16	-0.7685	-0.5257	-0.5473
17	-0.4031	-0.2062	0.4625
18	1.2304	0.7907	0.0762
19	1.0758	0.3764	0.2618
20	1.9928	-1.1419	0.4648

On trouvera donc dans SIMCA (*ibidem* p. 156) non pas -0.6596 mais $-0.6596\sqrt{19/20} = -0.643$. Cela ne change rien aux valeurs prédites. Nous avons vu que t_1 est la première coordonnée des lignes de l'analyse de co-inertie de X et Y. Au pas suivant, X est remplacé par les résidus des régressions simples de X sur t_1. Ce tableau est noté X_1 et on recommence l'analyse de co-inertie de Y avec X_1. t_2 est alors la première coordonnée normalisée des lignes de X_1. Au pas suivant X_1 est remplacé par les résidus des régressions simples de X_1 sur t_2. Ce tableau est noté X_2 et on recommence l'analyse de co-inertie de Y avec X_2 pour obtenir t_3.

La régression PLS est progressive. Au pas 1, on utilise une régression simple de chaque variable sur t_1. Au pas 2, on fait de même sous contrainte que le nouveau régresseur est non corrélées au précédent. Le nouveau modèle est évidemment moins bon que le précédent mais la somme des deux premiers est évidemment meilleur que le premier. Au total, on fabrique cependant un modèle linéaire de chaque variable

dépendante sur l'ensemble des explicatives. Nous avons vu que si toutes les composantes sont significatives on retrouve la MLR.

On peut s'intéresser au modèle final ou à la constitution progressive de ce modèle final. Le fichier :

A.PLSgen2w1

	1	2	3
1	-0.0773	-0.1380	-0.0603
2	-0.4995	-0.5250	-0.1559
3	-0.1323	-0.0855	-0.0072

contient les coefficients de régression (explicatives et les explicatives sont normalisées) du type :

$$y_k^* = \sum_{j=1}^{p} w_h x_h^*$$

Dans le tableau les explicatives sont en lignes et les dépendantes en colonne, ce qui se lit (*ibidem* p.162)

Tractions* = -.077 Poids* -0.499 Tour de taille* - 0.132 Pouls*

Le fichier :

A.PLSgen2w2

	1	2	3
1	47.0375	612.7674	183.9130
2	-0.0165	-0.3497	-0.1253
3	-0.8246	-10.2576	-2.4964
4	-0.0970	-0.7422	-0.0510

contient les modèle sur les données brutes (*ibidem* p. 162) :

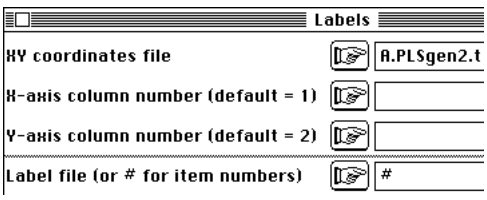
Tractions = 47.02 - 0.0166 Poids - 0.824 Tour de taille - 0.097 Pouls

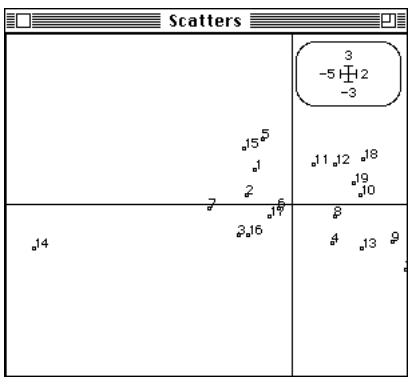
Les composantes t_1 et t_2 sont des systèmes de coordonnées obtenues par projection sur des axes de l'espace des lignes du tableau des explicatives. Ces coordonnées sont non corrélées comme pour des axes d'ACP. Mais les axes de projection ne forment pas une base orthonormale comme les axes d'ACP. La base des axes principaux de l'ACP est en effet la seule base orthonormée (rangée en colonne dans **U**) qui vérifie simultanément :

$\mathbf{U}^t\mathbf{X}^t\mathbf{D}\mathbf{X}\mathbf{U} = \Delta$ (orthogonalité des coordonnées)

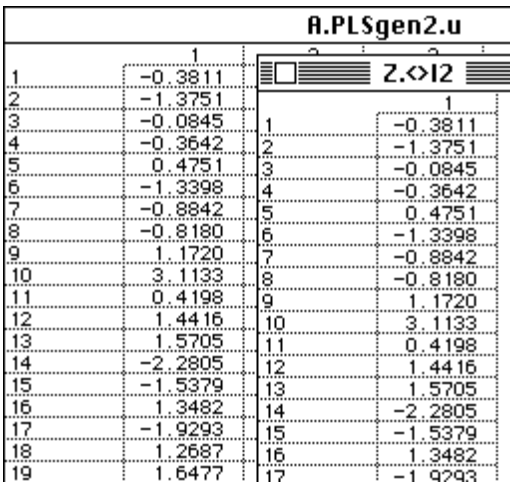
$\mathbf{U}^t\mathbf{U} = \mathbf{I}_p$ (orthogonalité des vecteurs)

Rien n'empêche de faire une représentation bivariée des composantes appelée plan des composantes t_1 et t_2 (*ibidem* p. 165) mais on devra se garder de prendre le dessin pour une projection euclidienne :



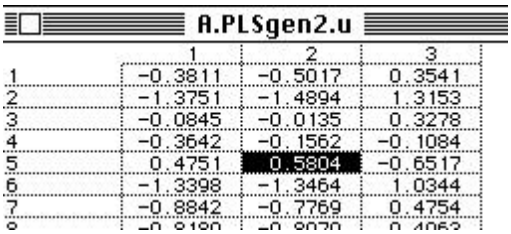


Comme il y a une analyse de co-inertie de rang 1 à chaque tour, les coordonnées qui sont dans t_1 correspondent à des coordonnées qui sont dans u_1^*, ce qu'on vérifie au premier tour en refaisant l'analyse de co-inertie entre les deux ACP normées (comme ci-dessus) :



	1
1	-0.3811
2	-1.3751
3	-0.0845
4	-0.3642
5	0.4751
6	-1.3398
7	-0.8842
8	-0.8180
9	1.1720
10	3.1133
11	0.4198
12	1.4416
13	1.5705
14	-2.2805
15	-1.5379
16	1.3482
17	-1.9293
18	1.2687
19	1.6477

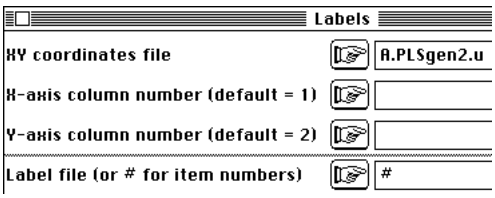
Le fichier :

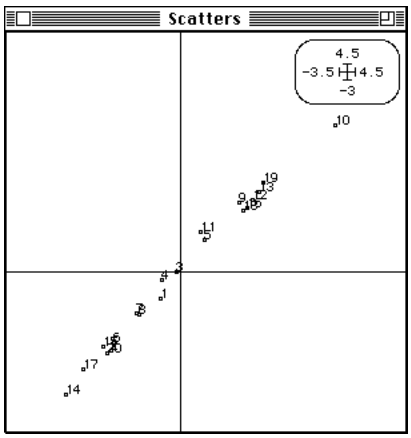


	1	2	3
1	-0.3811	-0.5017	0.3541
2	-1.3751	-1.4894	1.3153
3	-0.0845	-0.0135	0.3278
4	-0.3642	-0.1562	-0.1084
5	0.4751	0.5804	-0.6517
6	-1.3398	-1.3464	1.0344
7	-0.8842	-0.7769	0.4754

contient les composantes u_h^* à la constante déjà signalée près, c'est-à-dire qu'au lieu de -.5804 on trouve dans SIMCA (*ibidem* p. 157) la valeur $-0.5804\sqrt{19/20} = -0.566$.

La représentation bivariée associée est appelée plan des composantes u_1^* et u_2^* (*ibidem* 166) :





On obtiendra aisément par Scatters sur le fichier A.PLS2.ut qui contient en colonnes t_1, u_1^*, t_2, u_2^*, ... ou par Curves sur les fichiers A.PLS2.u et A.PLS2.t le graphique des composantes t_1 et u_1^* (*ibidem* p. 169) et celui des composantes t_2 et u_2^* (p. 170).

Le fichier :

A.PLSgen2.X+co

	1	2	3
1	-0.9476	-0.0128	-0.3191
2	-0.9620	-0.2349	0.1391
3	0.5108	-0.7901	-0.3389

contient les corrélations entre les explicatives (en lignes) et les composantes t_k (en colonnes) (*ibidem* p. 157), tandis que le fichier :

A.PLSgen2.Y+co

	1	2	3
1	0.4862	0.2229	-0.2314
2	0.5921	0.1927	-0.2209
3	0.2035	0.0431	0.1032

contient les corrélations entre les dépendantes (en lignes) et les composantes t_k (en colonnes). Ces valeurs sont réunies pour des représentations graphiques dans le fichier A.PLSgen2.XY+co.

Les composantes t_k servent dans deux types de régression, d'une part pour les variables à expliquer, d'autre part pour les explicatives. Dans le premier cas on construit le modèle final et dans le second cas on utilise le résidu au pas suivant. Dans les deux cas on dispose du pourcentage de variance expliquée par la composante sur les dépendantes (Fo) et pour les explicatives (Eo). Le listing donne ces valeurs :

```
|---------------|*** Eo ***|---------------------|
|Step| Variance | Explained|  Ratio   | Exp. Sum |
|   1| 1.000e+00| 6.948e-01| 6.948e-01| 6.948e-01|
|   2| 3.052e-01| 2.265e-01| 7.422e-01| 9.213e-01|


|---------------|*** Fo ***|---------------------|
|Step| Variance | Explained|  Ratio   | Exp. Sum |
|   1| 1.000e+00| 2.094e-01| 2.094e-01| 2.094e-01|
|   2| 7.906e-01| 2.955e-02| 3.737e-02| 2.390e-01|
```

Ces quantités sont appelées respectivement redondance de X et redondance de Y (*ibidem* p. 158). Comme toutes les variances initiales sont égales à l'unité, ces pourcentages de variance expliquée sont des moyennes de carrés de corrélation multiple par variable (*ibidem* p. 159) :

```
|---------------| Fo Col:  1|---------------------|
|Step| Variance | Explained |  Ratio   | Exp. Sum |
|   1| 1.000e+00|  2.363e-01| 2.363e-01| 2.363e-01|
|   2| 7.637e-01|  4.966e-02| 6.503e-02| 2.860e-01|


|---------------| Fo Col:  2|---------------------|
|Step| Variance | Explained |  Ratio   | Exp. Sum |
|   1| 1.000e+00|  3.506e-01| 3.506e-01| 3.506e-01|
|   2| 6.494e-01|  3.712e-02| 5.716e-02| 3.877e-01|


|---------------| Fo Col:  3|---------------------|
|Step| Variance | Explained |  Ratio   | Exp. Sum |
|   1| 1.000e+00|  4.140e-02| 4.140e-02| 4.140e-02|
|   2| 9.586e-01|  1.854e-03| 1.934e-03| 4.325e-02|
```

La variable 3 n'est pas prédictible par cette méthode.

On trouvera encore les vecteurs w_k, c_k et c_k^* respectivement dans les fichiers :

A.PLSgen2.w

	1	2	3
1	-0.5899	0.3635	-0.7467
2	-0.7713	-0.6902	0.6408
3	0.2389	-0.6257	-0.1784

A.PLSgen2.r

	1	2	3
1	0.3416	0.3419	-0.5967
2	0.4161	0.2956	-0.5694
3	0.1430	0.0661	0.2660

A.PLSgen2.c

	1	2	3
1	0.6133	0.7485	-0.6885
2	0.7470	0.6471	-0.6571
3	0.2567	0.1446	0.3070

(voir *ibidem* p. 155).

Une partie de ces aides à l'interprétation sont exploratoires et on pourra souvent préférer l'analyse de co-inertie simple à la régression PLS. Pour des questions proprement de prévisions la régression PLS l'emporte presque toujours sur l'ACPVI comme la PLS1 l'emporte presque toujours sur la MLR. Les régressions sur composantes sont dans la plupart des cas très proches. L'extension hors du champ des variables normalisées est essentielle dans le domaine écologique : c'est l'objectif mis en perspective par l'ouvrage de M. Tenenhaus.

Références

[1] Carrel, G. (1986) *Caractérisation physico-chimique du Haut-Rhône français et de ses annexes : incidences sur la croissance des populations d'alevins*. Thèse de doctorat. Université Lyon 1. 1-186.

[2] Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology : 24, 417-441 , 498-520.

[3] Rojas-Camousseight, F. (1985) *Etudes préliminaires sur l'utilisation des Trichoptères adultes comme descripteurs écologiques*. Thèse de doctorat, Université Lyon 1. 1-215.

Usseglio-Polatera, P. & Auda, Y. (1987) Influence des facteurs météorologiques sur les résultats de piégeage lumineux. *Annales de Limnologie* : 23, 1, 65-79.

[4] Tucker, L.R. . (1958) An inter-battery method of factor analysis. *Psychometrika* : 23, 2, 111-136.

Voir Chessel, D. & Mercier, P. (1993) Couplage de triplets statistiques et liaisons espèces-environnement. In : *Biométrie et Environnement*. Lebreton, J.D. & Asselain, B. (Eds.) Masson, Paris. 15-44.

[5] Tenenhaus, M. (1997) *Association et prédiction en analyse des données. De l'analyse canonique à la régression PLS*. Version provisoire. Groupe HEC - 78351 Jouy-en-Josas. 1-217.

[6] Tenenhaus, M. (1993) *La Régression PLS*. D 1614K93 Document Groupe HEC. Groupe HEC. 1-25.

Tenenhaus, M., Gauchi, J.P. & Ménardo, C. (1995) Régression PLS et applications. *Revue de Statistique Appliquée* : 43, 7-63.

[7] SIMCA. (1991) Soft Independent Modeling of Class Analogy. Version 4.3R. Umetri AB Box 1456, S - 90124 Umea.

SIMCA-P for Windows (1996) Graphical Software for Multivariate Process Modeling. Umetri AB Box 7960, S - 907 19 Umea, Sweden.