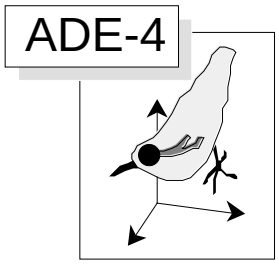


Entre ACP et ACM : l'analyse de Hill et Smith

Résumé

Les données écologiques sont fréquemment structurées en groupes de relevés en particulier dans le temps et dans l'espace. Les données morphométriques peuvent être aussi structurées en groupe de descripteurs. Ces descripteurs peuvent être qualitatifs ou quantitatifs. On les réunit par type pour des critères techniques bien que biologiquement une classe signifiante de descripteurs puissent comprendre des variables de plusieurs types. A partir d'un problème difficile d'analyse de données génétiques, on décrit l'usage de l'analyse de Hill & Smith (1976, Principal component analysis of taxonomic data with multi-state discrete characters. *Taxon* : 25, 249-255) qui permet de mélanger variables qualitatives et quantitatives.

Plan

D. Chessel & J.M. Lascaux

1 — Données traitées : approche exploratoire

Dans 10 cours d'eau de quatre bassins hydrographiques, on a récolté 179 truites (*Salmo trutta*) qu'on peut classer en 7 groupes génétiques par le nombre d'allèles méditerranéens. La variable génétique prend les modalités 1 (0 allèle méditerranéen, homozygotes atlantiques dites modernes), 2 à 6 (respectivement 1 à 5 allèles méditerranéens) et 7 (6 allèles méditerranéens, homozygotes méditerranéens dites ancestrales). Cette définition est déjà caractéristique des problèmes abordés ici : la variable génétique est-elle quantitative (nombre d'allèles) ou qualitative (catégories définies par le nombre d'allèles).

Chaque individu a fait l'objet d'un enregistrement multivarié de 4 types de descripteurs, groupant respectivement les variables méristiques, les variables morphométriques, les caractéristiques de la ponctuation et les paramètres qualitatifs décrivant les traits ornementaux du poisson.

Dans les quatre tableaux de variables, on cherche à caractériser les différences qui peuvent se manifester entre les deux lignées méditerranéenne (truites sauvages) et atlantique (truites déversées) dans la région étudiée.

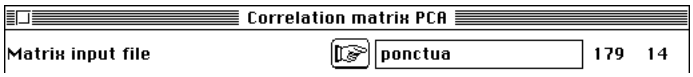
Cette première partie explore à travers divers modules d'ADE-4 les questions que posent la manipulation de quatre groupes de descripteurs de nature variée dans une analyse visant la discrimination entre groupes.

1.1 — Variables de ponctuation

Les variables de ponctuation sont au nombre de 14 :

PRAD	Nombre de points rouges avant l'aplomb de la Dorsale
PNAD	Nombre de points noirs avant l'aplomb de la Dorsale
PRAA	Nombre de points rouges après l'aplomb de l'Anale
PNAA	Nombre de points noirs après l'aplomb de l'Anale
PRDA	Nombre de points rouges entre aplomb Dorsale et aplomb Anale
PNDA	Nombre de points noirs entre aplomb Dorsale et aplomb Anale
PRLI	Nombre de points rouges sur la ligne latérale
PRINF	Nombre de points rouges au dessous de la ligne latérale
PNINF	Nombre de points noirs au dessous de la ligne latérale
PRSUP	Nombre de points rouges au dessus de la ligne latérale
PNSUP	Nombre de points noirs au dessus de la ligne latérale
PNO	Nombre de points noirs sur l'opercule
PRD	Nombre de points rouges sur la Dorsale
PND	Nombre de points noirs sur la Dorsale

Les observations (après transformation de toutes les variables par y = Log(x+1)) sont dans un fichier ponctua (179-14) dont l'analyse en composantes principales est un cas d'école :




```
Classical Principal Component Analysis (Hotteling 1933)
Input file: ponctua
---- Row weights:
File ponctua.cnpl contains the row weights
It has 179 rows and 1 column
Each row has 5.5866e-03 weight (Sum = 1)
```

```
---- Column weights:
File ponctua.cnpc contains the column weights
It has 179 rows and 1 column
Each column has unit weight (Sum = 14)
---- Table:
File ponctua.cnta contains the centred and normed table
Zero mean and unit variance for each column
It has 179 rows and 14 columns
File :ponctua.cnta
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-3.499e+00| 1.902e+00|
|   2|-2.196e+00| 1.977e+00|
|   3|-4.128e+00| 2.434e+00|
|   4|-2.713e+00| 2.123e+00|
|   5|-4.810e+00| 2.257e+00|
|   6|-2.733e+00| 1.892e+00|
|   7|-5.971e+00| 1.236e+00|
|   8|-3.920e+00| 1.900e+00|
|   9|-2.250e+00| 1.785e+00|
|  10|-2.126e+00| 2.164e+00|
|  11|-3.377e+00| 1.970e+00|
|  12|-2.513e+00| 1.721e+00|
|  13|-1.529e+00| 1.591e+00|
|  14|-3.828e+00| 1.314e+00|
|----|----------|----------|
```

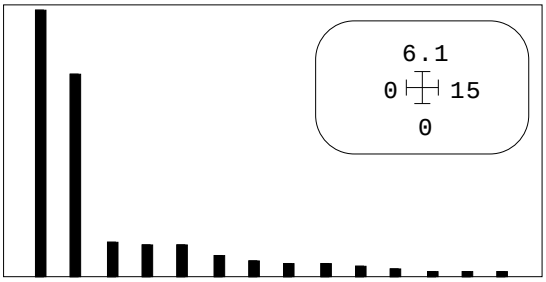
Certaines valeurs du tableau normalisé sont anormalement basses (en particulier -6 fois l'écart-type) qui laisse penser à la présence d'un individu ayant une ponctuation en points rouge anormalement basse.

```
------------------------------------------------------------------------------
[  1] 1000
[  2] -393 1000
[  3]  630  -72 1000
[  4] -329  773  -41 1000
[  5]  726  -56  810  -32 1000
[  6] -359  914  -35  870  -34 1000
[  7]  572  -31  633   23  680   -8 1000
[  8]  805 -193  840 -147  897 -162  555 1000
[  9] -451  921 -117  722 -105  855  -78 -220 1000
[ 10]  722 -166  767 -153  800 -138  382  741 -243 1000
[ 11] -296  913   12  895   20  974   49 -113  794  -85 1000
[ 12] -320  842   -8  676  -54  777   33 -145  809 -157  767 1000
[ 13]  417   79  470   33  466   93  352  485   27  399  107  161 1000
[ 14] -163  538  123  485   57  557   78  -27  432   53  577  627  133 1000
------------------------------------------------------------------------------
```

La matrice des corrélations donne des corrélations négatives entre couple de variables rouge-noir et des corrélations en général positive entre variables rouge-rouge ou noir-noir.

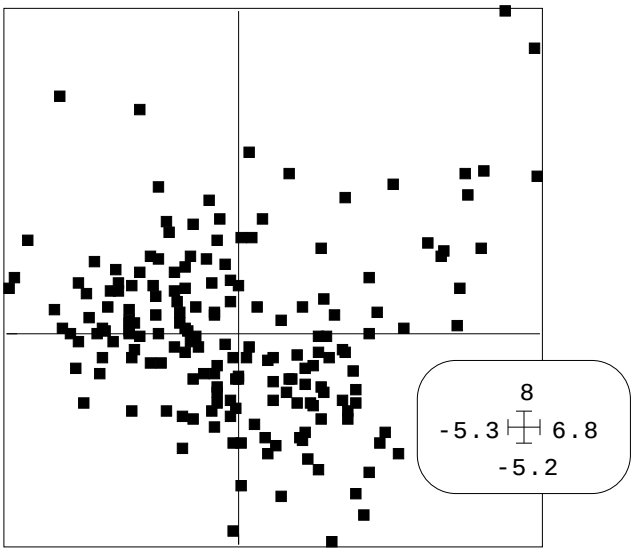
```
-------------------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: ponctua.cnta
--- Number of rows: 179, columns: 14
-----------------------
Total inertia:       14
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +6.0362E+00 +0.4312 +0.4312   |02   +4.5386E+00 +0.3242 +0.7553 |
03   +7.1274E-01 +0.0509 +0.8063   |04   +6.6982E-01 +0.0478 +0.8541 |
05   +6.2053E-01 +0.0443 +0.8984   |06   +4.0095E-01 +0.0286 +0.9271 |
07   +2.9319E-01 +0.0209 +0.9480   |08   +2.2580E-01 +0.0161 +0.9641 |
•••
File ponctua.cnvp contains the eigenvalues and relative inertia for each
axis
```

La répartition de l'inertie est particulièrement sympathique :

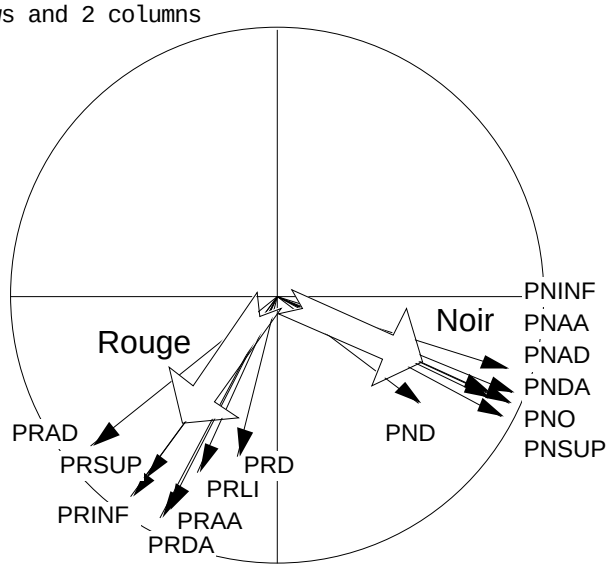


Le nuage des points est dans un plan, le reste de la variabilité est sans conteste un bruit aléatoire :

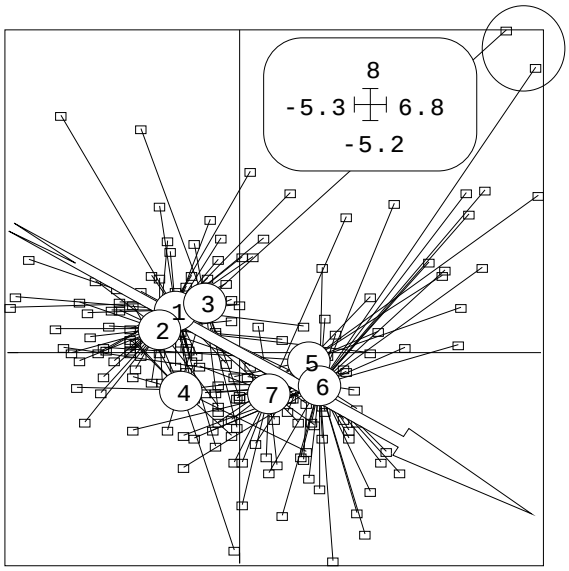
```
File ponctua.cnli contains the row scores
--- It has 179 rows and 2 columns
```



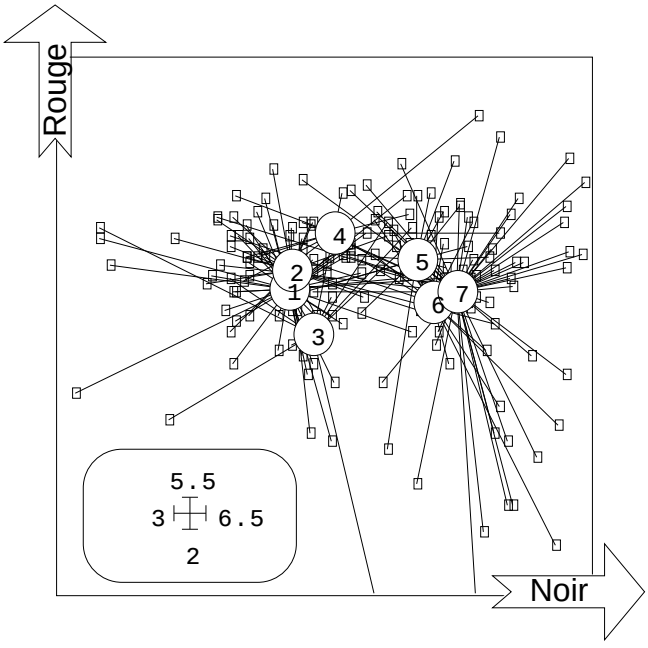

```
File ponctua.cnco contains the column scores
--- It has 14 rows and 2 columns
```



Le nuage des variables est donc aussi dans un plan et le cercle des corrélations est particulièrement signifiant. Le lien entre ponctuation et variable génétique s'impose naturellement :



Le nombre total de points noirs est fortement lié à la variable génétique. On peut cependant craindre que l'influence de la taille des poissons soit importante et perturbe un résultat en apparence satisfaisant. Nous pouvons seulement dire que la typologie des individus obtenus avec les 14 variables de ponctuation est bien résumée par deux variables, respectivement le nombre total de points rouges (1+3+5+7+8+10+13) et le nombre total de points noirs (2+4+6+9+11+12+14). En témoigne la représentation des données simples :

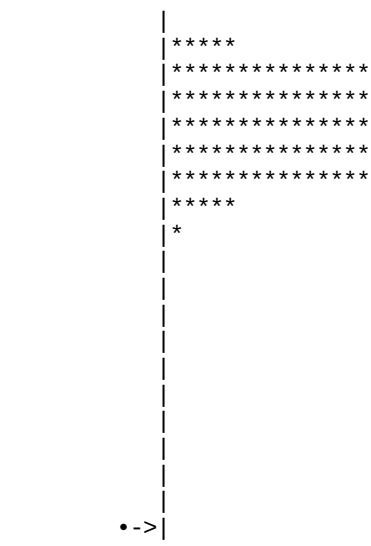


Si l'ACP est si claire, c'est en partie parce qu'une partie des corrélations provient d'un artefact. Les points sont comptés plusieurs fois et ce mode d'enregistrement augmente le poids de la densité des points au détriment de l'analyse du pattern de répartition. Une des propriétés fondamentales de l'analyse discriminante étant de donner des résultats invariants par une transformation linéaire des données, il est utile de voir le résultat obtenu.

On utilise Discrimin : Initialize/LinkPrep, Discrimin : Discriminant analysis/Test et Discrimin : Discriminant analysis/Run :

```
Categories defined by column 1 of file Groupe
```

```
Input statistical triplet: table ponctua.cnta - total inertia: 14.000000
between class inertia 3.412483 (ratio: 0.243749)
within class inertia 10.587517 (ratio: 0.756251)

number of random matching: 10000   Observed: 1.217239
Histogramm:  minimum = 0.259507, maximum = 1.217239
number of simulation X<Obs: 10000 (frequency: 1.000000)
number of simulation X>=Obs: 0 (frequency: 0.000000)
```



```
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +7.7863E-01 +0.6397 +0.6397   |02   +1.4848E-01 +0.1220 +0.7617 |
03   +1.1462E-01 +0.0942 +0.8558   |04   +8.1371E-02 +0.0668 +0.9227 |
```

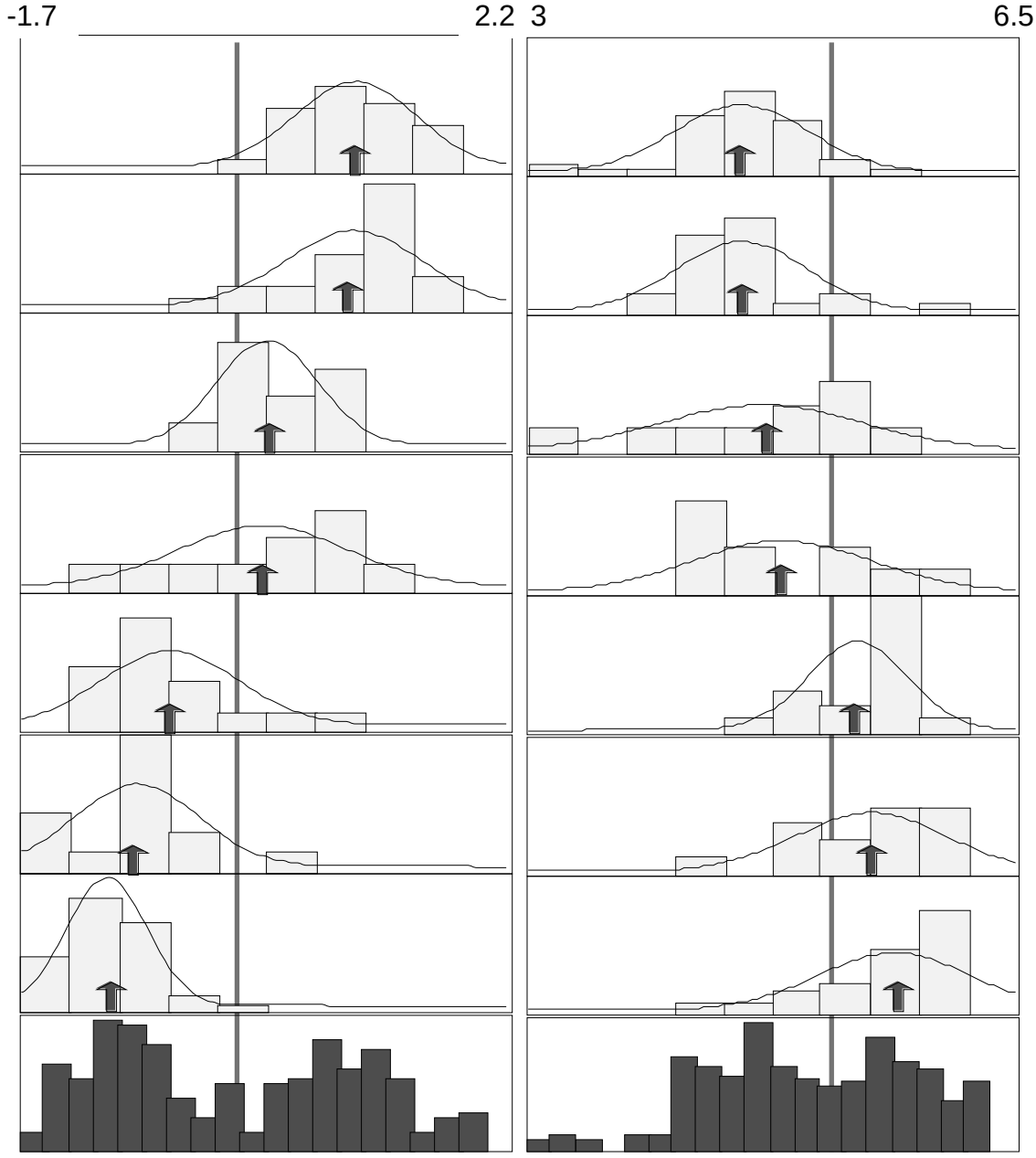
L'analyse est très hautement significative. Un seul facteur donnant 77% de variables expliquée est conservé. On compare ensuite poids des variables dans l'analyse canonique et corrélations variables observées / variable canonique (par MatAlg : Diagonal Inner product C=X'DY).

```
File A.difa contains coefficient of discriminant scores
It has 14 rows and 1 columns
```

Variable	Poids	Corrélation
PRAD	0.3899	0.3280
PNAD	0.1966	*-0.7619*
PRAA	0.1283	-0.0536
PNAA	-0.0969	*-0.5994*
PRDA	0.3217	0.0152
PNDA	-0.3026	*-0.7026*
PRLI	-0.3671	-0.1372
PRINF	-0.2094	0.1144
PNINF	-0.4110	*-0.8198*
PRSUP	-0.2769	0.1536
PNSUP	0.5389	*-0.6502*
PNO	-0.7927	*-0.9153*
PRD	-0.2990	-0.3507
PND	0.0988	*-0.4811*

Toutes les valeurs élevées des corrélations portent sur les dénombrements de points noirs mais on note les contradictions habituelles entre corrélation d'un signe et poids canonique de l'autre signe. La discrimination est optimisée par la variable canonique, c'est sa définition, sans qu'on sache si cette optimalité réalisée a une signification certaine.

```
File A.dili contains canonical row scores with unit norm
It has 179 rows and 1 columns

File :A.dili
----------------------Minimum/Maximum:
Col.:   1  Mini =   -1.69    Maxi =  2.1385
```



On observe à gauche la discrimination opérée par la variable canonique et à droite la discrimination opérée par la simple observation de la somme totale du nombre de points noirs (en Log). On préfère à gauche la bimodalité et à droite la simplicité du descripteur. Il n'y a pas contradiction.

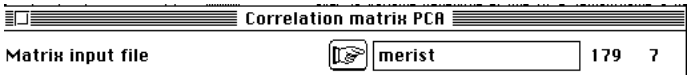
Au bilan de cette première approche, on retient une liaison certaine de ce groupe de variables avec la variable génétique et une ACP remarquable à deux gradients "coloration rouge" et "coloration noire" quasiment indépendants. La discrimination porte essentiellement sur le second.

1.2 — Variables méristiques

Les variables méristiques sont au nombre de 7 :

RD	Nombre de rayons à la dorsale
RA	Nombre de rayons à l'anale
RPELG	Nombre de rayons à la pelvienne gauche
RPECG	Nombre de rayons à la pectorale gauche
BR	Nombre de branchiospines (1°arc gauche)
VERT	Nombre de vertèbres
CAEC	Nombre de caeca pyloriques

Les observations sont dans un fichier merist (179-7) dont l'analyse en composantes principales est un cas d'école :




```
Classical Principal Component Analysis (Hotteling 1933)
Input file: merist
---- Row weights:
File merist.cnpl contains the row weights
It has 179 rows and 1 column
Each row has 5.5866e-03 weight (Sum = 1)
---- Column weights:
File merist.cnpc contains the column weights
It has 179 rows and 1 column
Each column has unit weight (Sum = 7)
---- Table:
File merist.cnta contains the centred and normed table
Zero mean and unit variance for each column
It has 179 rows and 7 columns
File :merist.cnta
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-3.243e+00| 1.705e+00|
|   2|-4.111e+00| 1.367e+00|
|   3|-5.087e+00| 5.031e+00|
|   4|-3.698e+00| 1.474e+00|
|   5|-2.579e+00| 2.758e+00|
|   6|-1.884e+00| 1.863e+00|
|   7|-2.065e+00| 2.717e+00|
|----|----------|----------|
```

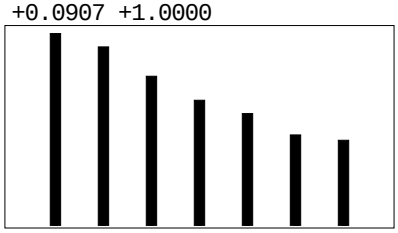
Il y a peut-être un doute sur des valeurs extrêmes dans la variable 5.

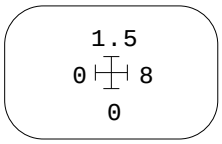
```
----------------------- Correlation matrix -------------------
[  1] 1000
[  2]  227 1000
[  3]   48 -116 1000
[  4]   62  166   90 1000
[  5] -161   50    3  237 1000
[  6]  127   65  -53   82  -24 1000
[  7]  -55 -149  153   41  147  -83 1000
------------------------------------------------------------
```

La matrice des corrélations contient des valeurs remarquablement faibles.

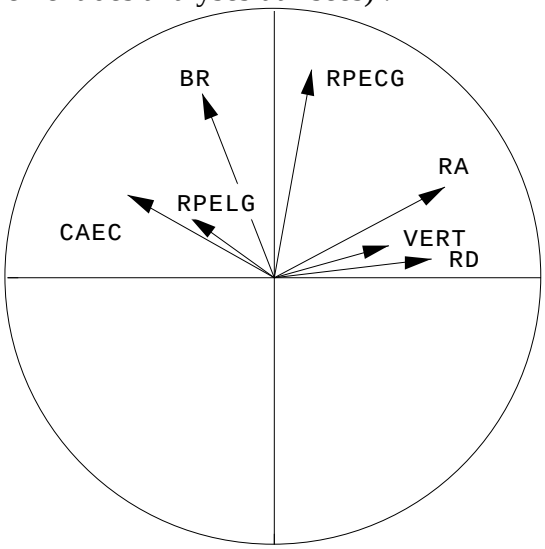
```
DiagoRC: General program for two diagonal inner product analysis
Input file: merist.cnta
--- Number of rows: 179, columns: 7
-----------------------
Total inertia:        7
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
```

```
01   +1.4518E+00 +0.2074 +0.2074   |02   +1.3403E+00 +0.1915 +0.3989 |
03   +1.1164E+00 +0.1595 +0.5584   |04   +9.4451E-01 +0.1349 +0.6933 |
05   +8.3484E-01 +0.1193 +0.8126   |06   +6.7703E-01 +0.0967 +0.9093 |
07   +6.3503E-01 +0.0907 +1.0000
```





Il n'y a aucune structure de corrélation inter-variables. Il n'y a rien à apprendre des cartes factorielles (sinon ce qu'on y voit quand il n'y a rien à voir, c'est-à-dire les contraintes de fonctionnement des analyses utilisées) :



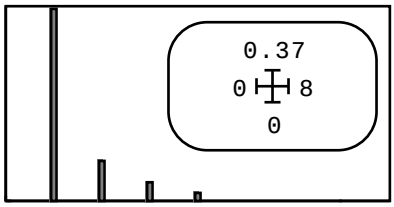
Les variables méristiques, quasiment indépendantes sont idéales pour la discrimination qui sera quasiment une discrimination sur variables orthogonales.

```
Between and Within-class inertia
Categories defined by column 1 of file Groupe
Input statistical triplet: table merist.cnta
total inertia: 7.000000
between class inertia 0.559337 (ratio: 0.079905)
within class inertia 6.440663 (ratio: 0.920095)
```

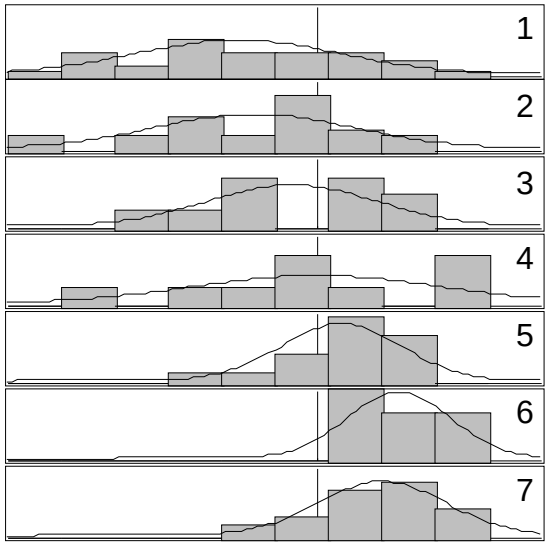
Il n'y a que 8% de variabilité inter-groupes. Mais cette valeur est hautement significative :

```
    |
    |***
    |*********
    |**********************
    |***********************************
    |*****************************************
    |************************************************
    |*****************************************
    |***********************************
    |************************
    |**************
    |********
    |***
    |**
    |
    |
    |
    |
    |
•->|
```

```
number of random matching: 10000   Observed: 0.508059
Histogramm:  minimum = 0.095134, maximum = 0.508059
number of simulation X<Obs: 10000 (frequency: 1.000000)
number of simulation X>=Obs: 0 (frequency: 0.000000)

Discriminant analysis
Categories defined by column 1 of file Groupe
Input statistical triplet: table merist.cnta
Number of rows: 179, columns: 7
total inertia (norm C- generalised inverse) = rank of the data matrix: 7.000000

between-class inertia (norm C-): 0.508059 (ratio: 0.072580)

Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +3.6616E-01 +0.7207 +0.7207   |02   +8.0425E-02 +0.1583 +0.8790 |
03   +3.5763E-02 +0.0704 +0.9494   |04   +1.8442E-02 +0.0363 +0.9857 |
05   +5.0636E-03 +0.0100 +0.9957   |06   +2.2084E-03 +0.0043 +1.0000 |
```



Il est prudent de ne considérer que la première variable canonique. Le fichier B.dili permet de représenter cette variable par groupe :



On retrouve un gradient très voisin du précédent. Le fichier B.difa donne les coefficients constitutif de cette variable canonique soit :

0.0449, 0.2289,-0.0189, 0.0245, 0.1163, 0.0100 et **-0.9377**

Le dernier l'emporte fortement sur les autres. Les corrélations variables - variable canonique confirme :

0.1315, 0.3916, -0.1846, 0.0532, -0.0119, 0.1085 et **-0.9598**

Le nombre de caeca pyloriques est la seule variable du groupe qui est en liaison avec la génétique.

On confirme avec l'analyse de variance (Discrimin : Anova1-FF) :

variable 1 from merist versus variable 1 from Groupe

Source	SS	d.f.	MS	F	Proba
Between	3.634	6	0.6057	1.676	0.1286
Within	62.16	172	0.3614		
Total	65.8	178			

variable 2 from merist versus variable 1 from Groupe

Source	SS	d.f.	MS	F	Proba
Between	3.904	6	0.6507	2.248	0.04065
Within	49.78	172	0.2894		
Total	53.69	178			

variable 3 from merist versus variable 1 from Groupe

Source	SS	d.f.	MS	F	Proba
Between	0.295	6	0.04917	1.262	0.2766
Within	6.699	172	0.03895		
Total	6.994	178			

variable 4 from merist versus variable 1 from Groupe

Source	SS	d.f.	MS	F	Proba
Between	0.3809	6	0.06348	0.1824	0.9798
Within	59.84	172	0.3479		
Total	60.22	178			

variable 5 from merist versus variable 1 from Groupe

Source	SS	d.f.	MS	F	Proba
Between	5.17	6	0.8616	0.6705	0.6756
Within	221	172	1.285		
Total	226.2	178			

variable 6 from merist versus variable 1 from Groupe

Source	SS	d.f.	MS	F	Proba
Between	2.537	6	0.4229	0.6483	0.6935
Within	112.2	172	0.6523		
Total	114.7	178			

variable 7 from merist versus variable 1 from Groupe

Source	SS	d.f.	MS	F	Proba
Between	6095	6	1016	14.63	0
Within	1.194E+04	172	69.43		
Total	1.804E+04	178			

De ce groupe de 7 variables, dépourvu de corrélation entre paramètres, un seul descripteur est très lié à la partition génétique. Les autres seront désormais exclus de la discussion.

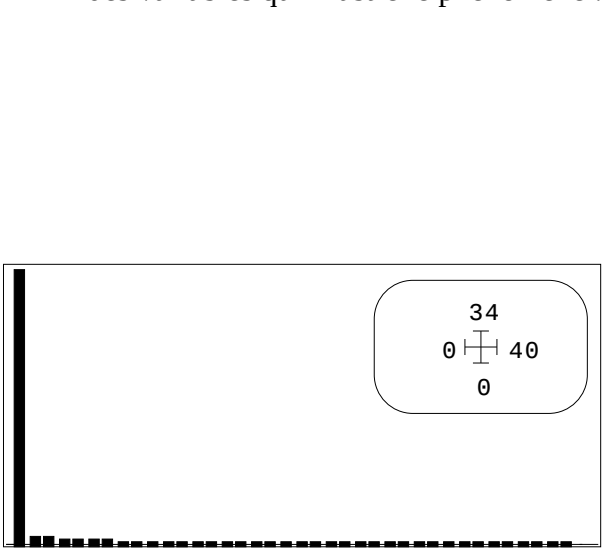
1.3 — Variables morphométriques

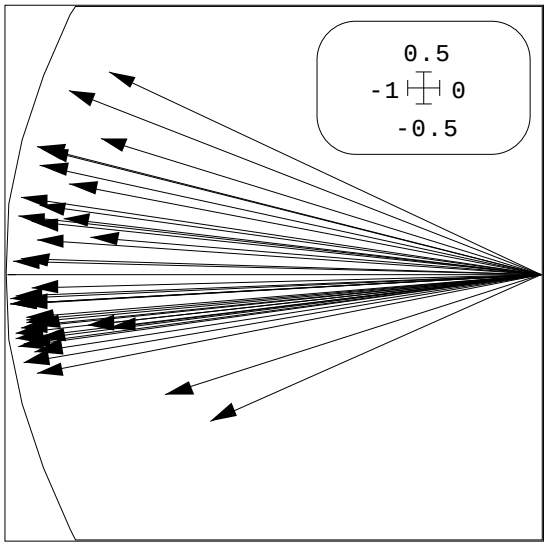
Les variables méristiques sont au nombre de 39 :

TPR	Diamètre des plus gros points rouges (dans la région de la Dorsale)
TPN	Diamètre des plus gros points noirs (dans la région de la Dorsale)
LS	Longueur standard

MD	Distance bout du museau - insertion de la dorsale
MAD	Distance bout du museau - insertion de l'adipeuse
MAN	Distance bout du museau - insertion de l'anale
MPEL	Distance bout du museau - insertion de la pelvienne
MPEC	Distance bout du museau - insertion de la pectorale
DAD	Distance insertion de la dorsale - insertion de l'adipeuse
DC	Distance insertion de la dorsale - départ de la caudale
DAN	Distance insertion de la dorsale - insertion de l'anale
DPEL	Distance insertion de la dorsale - insertion de la pelvienne
DPEC	Distance insertion de la dorsale - insertion de la pectorale
ADC	Distance insertion de l'adipeuse - départ de la caudale
ADAN	Distance insertion de l'adipeuse - insertion de l'anale
ADPEL	Distance insertion de l'adipeuse - insertion de la pelvienne
ADPEC	Distance insertion de l'adipeuse - insertion de la pectorale
PECPEL	Distance insertion de la pectorale - insertion de la pelvienne
PECAN	Distance insertion de la pectorale - insertion de l'anale
PECC	Distance insertion de la pectorale - départ de la caudale
PELAN	Distance insertion de la pelvienne - insertion de l'anale
PELC	Distance insertion de la pelvienne - départ de la caudale
ANC	Distance insertion de l'anale - départ de la caudale
LPRO	Longueur préorbitale
DO	Diamètre de l'orbite
LPOO	Longueur postorbitale
LTET	Longueur de la tête
HTET	Hauteur de la tête (en passant par au milieu de l'orbite)
LMAX	Longueur de la mâchoire supérieure
LAD	Longueur de l'adipeuse
LD	Longueur de la dorsale
HD	Hauteur de la dorsale
LC	Longueur de la caudale
LAN	Longueur de l'anale
HAN	Hauteur de l'anale
LPELG	Longueur de la pelvienne gauche
LPECG	Longueur de la pectorale gauche
HPED	Hauteur du corps au niveau du pédoncule caudal
ETET	Largeur de la tête (au niveau des orbites)

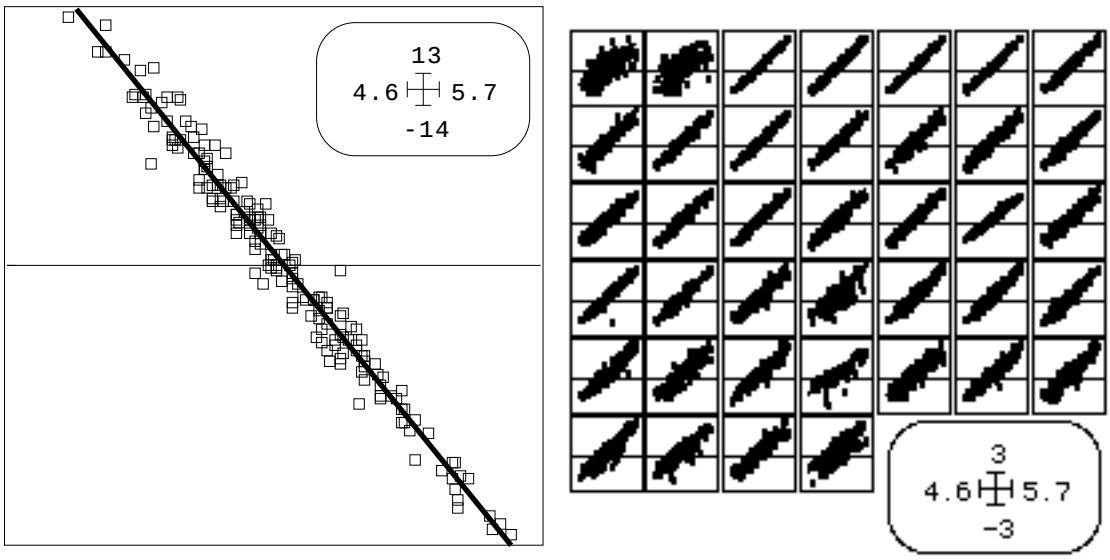
Les logarithmes des valeurs observées sont dans un fichier morpholog (179-39) dont l'ACP est encore un cas d'école. 86% d'inertie sont représentées sur le premier axe, ce qui fait disparaître toute nuance chez les autres. On reconnaît l'effet taille sur la carte 1-2 des variables qui illustre le phénomène :



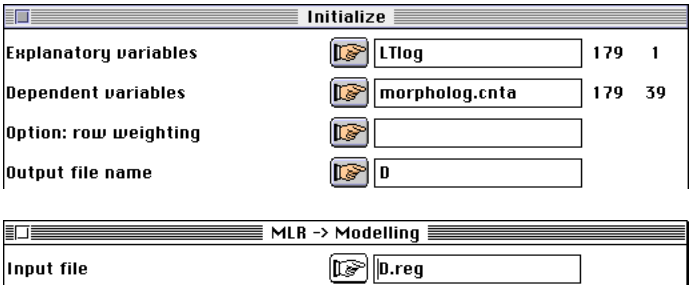


Il faut débarrasser les données de l'effet taille, d'autant plus que la majorité des variables présente des différences de moyennes hyper-significatives entre groupe génétique. Le fait dérive des contraintes d'échantillonnage, des conditions écologiques et de la dynamique des populations : il ne saurait être un parasité toléré vus les objectifs de l'étude.

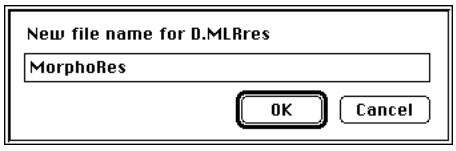
Le double centrage des données en logarithmes pourrait être une solution. Cependant les variables de ponctuation pourraient également être concernée par l'effet taille et il vaut mieux avoir un traitement global. D'autant plus que la liaison entre le facteur 1 de l'ACP qui précède (en ordonnée, à gauche) et la longueur totale des poissons est très forte, conséquences des liaisons très fortes entre les variables de départ (en ordonnée, à droite) et la même longueur totale (en log, en abscisse dans toutes les fenêtres) :



Le plus simple est de corriger toutes les variables morphométrique de l'effet taille par une régression simple dans LinearReg :

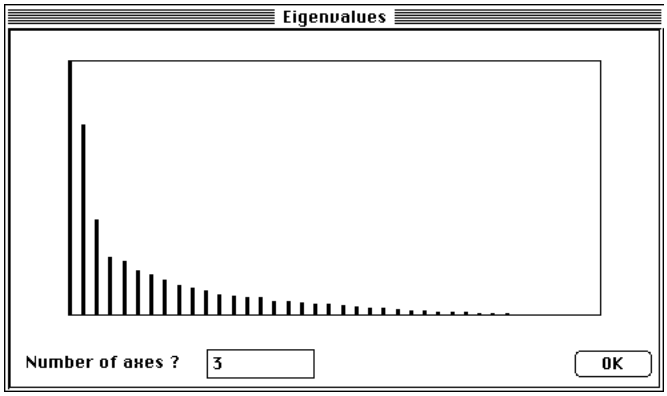


Les variables sont expliquée à hauteur de 70 à 99%. Changer le nom du tableau des résidus :

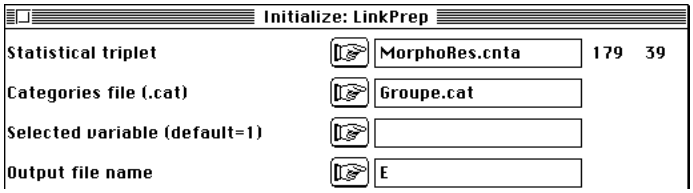


Faire l'ACP normée de ce fichier :

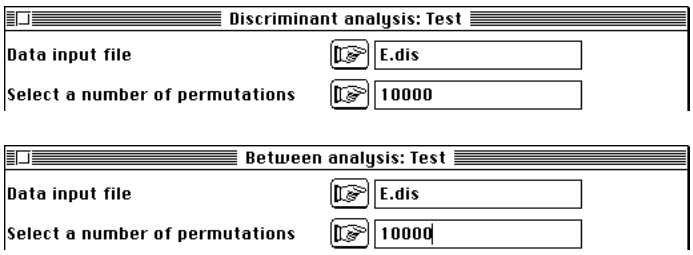
On obtient une structure forte des corrélations partielles (sachant l'effet de la taille) :



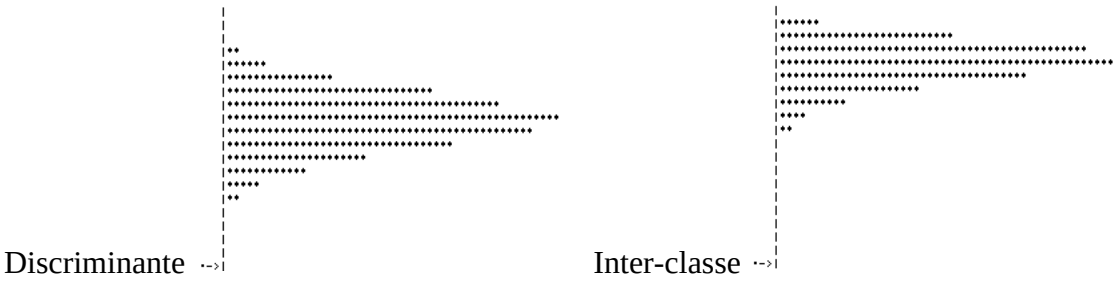
Initier l'inter-groupes associée (Discrimin : Initialize/LinkPrep) :



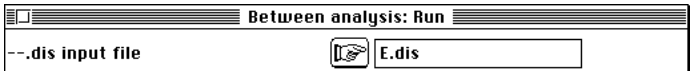
Tester l'analyse discriminante puis l'inter-classe :

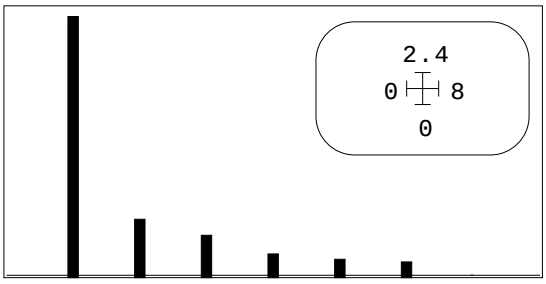


On préfère la seconde à la première, vus le grand nombre de variable et la signification supérieure qui correspond à une séparation des groupes associée à une structure des variables :

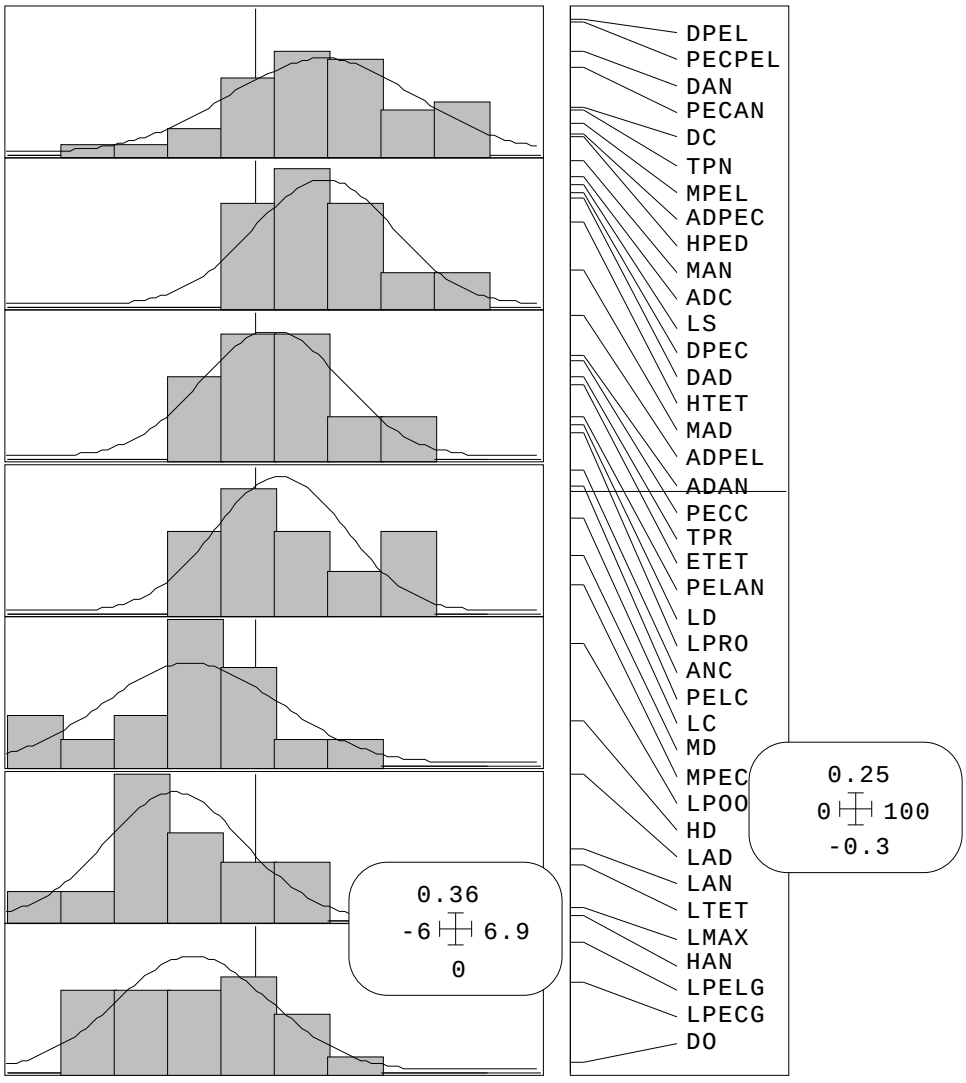


Exécuter l'analyse inter-classe :





Toute l'information est concentrée à nouveau dans une seule variable canonique :



Il devient surprenant de retrouver un modèle très voisin des précédents. La participation des variables est décrite par les poids canoniques.

1.4 — Paramètres ornementaux

Les variables photographiques sont au nombre de 16 regroupant 36 modalités :

TâPéri	Nombre de taches péri operculaires	1a : 2-3
		1b : 4-5
		1c : 6-7-8
OcPR	Ocelles autour des points rouges	2a : pas d'ocelles ou très peu marquées

2b : peu marquées
2c : marquées

OcPN Ocelles autour des points noirs 3a : pas d'ocelles ou très peu marquées
3b : peu marquées
3c : marquées

MaJu Marques Juvéniles 4a : absence
4b : présence

TâOp Tâche operculaire 5a : absence
5b : présence

PNTêt Points noirs sur la tête 6a : absence
6b : Présence

FrD Frange de la Dorsale 7a : pas de frange
7b : frange blanche ou blanche et noire

PtsAd Points sur l'adipeuse 8a : absence
8b : Présence

FrAd Frange de l'Adipeuse 9a : pas de frange
9b : frange plus ou moins rouge

FrAn Frange de l'Anale Aa : pas de frange ou frange blanche
Ab : frange blanche et noire

FrPel Frange des Pelviennes Ba : pas de frange
Bb : frange blanche
Bc : frange blanche et noire

FrC Frange de la Caudale Ca : pas de frange
Cb : frange plus ou moins rouge

PtsDos Points sur le dos du poisson Da : absence
Db : présence

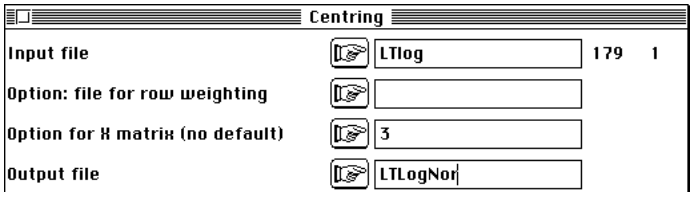
CouFI Couleurs des flancs du poisson Ea : brun jaune
Eb : gris à blanc argenté

ConPN Contour des points noirs du flanc Fa : net
Fb : flou

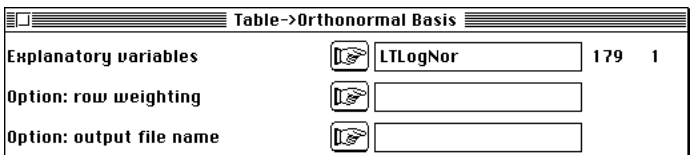
Zéb Zébrures sur le flanc du poisson Ga : absence
Gb : présence

Elles forment un fichier de variables qualitatives Photo (179-16). Les analyses de variances (Discrimin : Anova1-FF) montre que plusieurs de ces variables sont très fortement liées à la taille du poisson. On initie une analyse des correspondances multiples (MCA : Multiple Correspondence Analysis) :

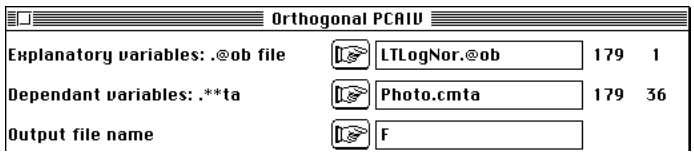
La variable longueur totale (en Log) est normalisée (Bin->Bin : Centring) :

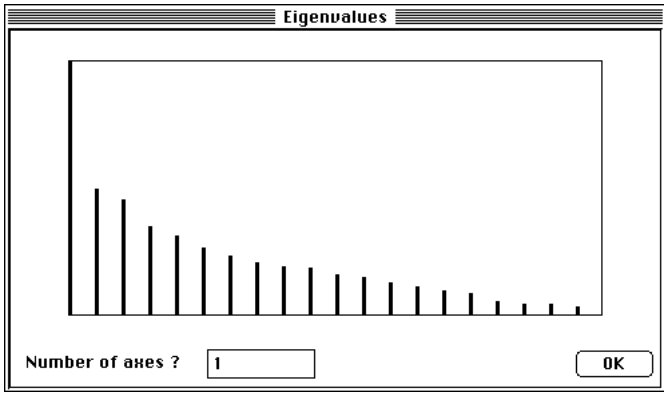


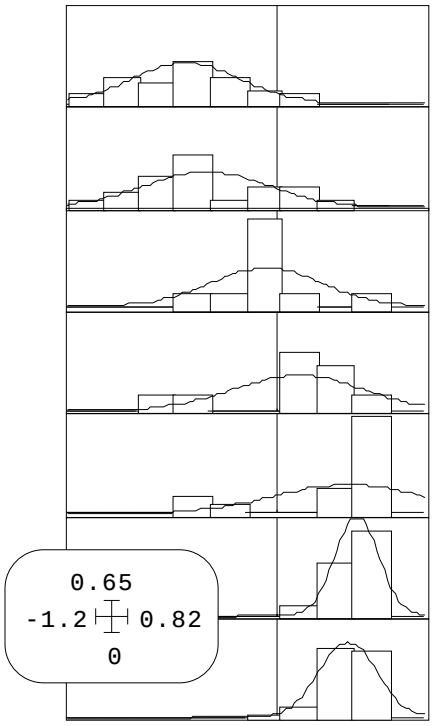
On définit un sous-espace de dimension 1 (Projectors : Table->Orthonormal Basis) :



On exécute l'ACPVI orthogonale associée (Projectors : Orthogonal PCAIV) :





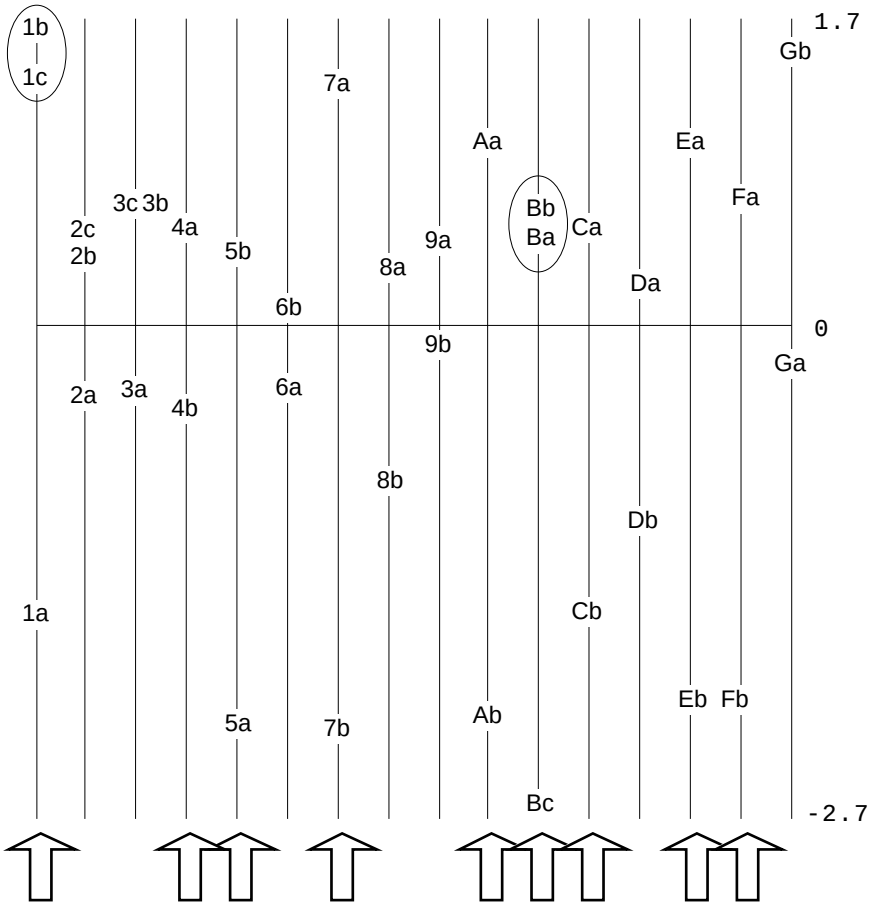


On obtient une structure unidimensionnelle qui reproduit directement le modèle de liaison avec la variable génétique. On a simplement fait ici une analyse des correspondances multiples sous contraintes de non corrélation des coordonnées des individus avec la taille. Toute la variabilité organisée du tableau de variables photographiques est d'essence génétique.

Dans deux des quatre tableaux et vraisemblablement dans un troisième (ponctuation) l'effet taille est une perturbation forte de l'analyse de l'influence de la variable génétique. Les quatre tableaux conduisent au même modèle de cette influence. La variable génétique est quantitative et le lien ne se fait avec le numéro de la classe qu'à travers le nombre d'allèles qui définit ce numéro. Enfin les quatre tableaux donne quatre types d'intervention des variables (une seule, une somme d'une partie, un grand nombre, un codage des modalités).

```
File ABB contains crossed Burt's matrix X'DY from
coding matrix X and coding matrix Y.
It has 7 rows (categories) and 36 columns (categories)
Access for X: file Groupe.cat
Access for Y: file Photo.cat
File ABB.bli contains row indicator (number of modalities for each variable)
It has 1 rows (categories) and 1 column
File ABB.bco contains column indicator (number of modalities for each variable)
It has 16 rows (categories) and 1 column
-----------------------------------------------------------------
        Khi2 tests on Burt's crossed table ABB
-----------------------------------------------------------------
Var   1 versus var   1:Khi2 =  1.3311e+02 ddl =   12 Proba =  0.000000
```

```
Var   1 versus var   2:Khi2 =  1.8677e+01 ddl =   12 Proba =  0.096178
Var   1 versus var   3:Khi2 =  1.9303e+01 ddl =   12 Proba =  0.081098
Var   1 versus var   4:Khi2 =  2.7234e+01 ddl =    6 Proba =  0.000159
Var   1 versus var   5:Khi2 =  4.0340e+01 ddl =    6 Proba =  0.000001
Var   1 versus var   6:Khi2 =  2.2319e+01 ddl =    6 Proba =  0.001154
Var   1 versus var   7:Khi2 =  1.1075e+02 ddl =    6 Proba =  0.000000
Var   1 versus var   8:Khi2 =  5.9614e+00 ddl =    6 Proba =  0.428068
Var   1 versus var   9:Khi2 =  5.4373e+00 ddl =    6 Proba =  0.490170
Var   1 versus var  10:Khi2 =  7.2969e+01 ddl =    6 Proba =  0.000000
Var   1 versus var  11:Khi2 =  6.2563e+01 ddl =   12 Proba =  0.000000
Var   1 versus var  12:Khi2 =  3.4639e+01 ddl =    6 Proba =  0.000008
Var   1 versus var  13:Khi2 =  8.0145e+00 ddl =    6 Proba =  0.236118
Var   1 versus var  14:Khi2 =  8.7428e+01 ddl =    6 Proba =  0.000000
Var   1 versus var  15:Khi2 =  6.0301e+01 ddl =    6 Proba =  0.000000
Var   1 versus var  16:Khi2 =  1.4624e+01 ddl =    6 Proba =  0.023340
---------------------------------------------------------------
```



La partie exploratoire a utilisé des ACP, des régressions, des inter-classes, des discriminantes et une ACP sur variables instrumentales. On sent la nécessité d'intégrer les variables en blocs, non par types techniques (qualitatif contre quantitatif) mais par type biologique (description de la robe contre description de la forme). Pour ce faire, nous montrons dans la seconde partie comment la liaison entre deux variables se mesure facilement de manière unique et dans la troisième partie comment qualitatives et quantitatives peuvent se mélanger dans une même analyse à un tableau.

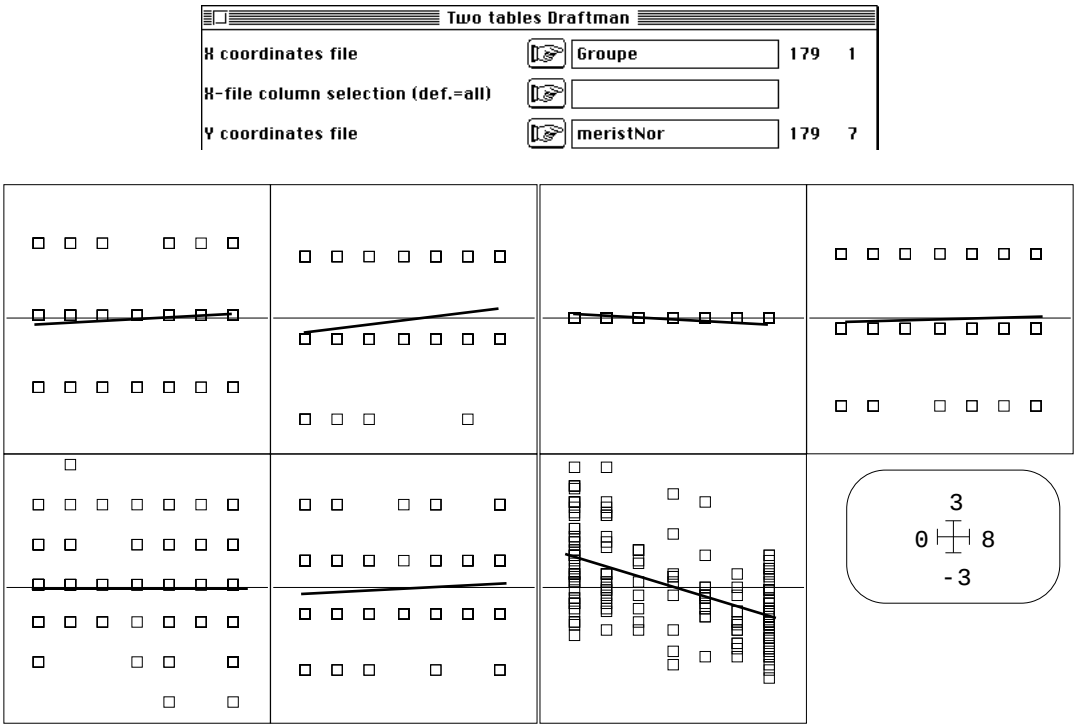
2 — Le lien entre deux variables

Ce qui est simple est le plus efficace. On veut dresser le tableau du rôle respectif de la taille et de la génétique pour les différentes variables. On veut mesurer la liaison d'une manière homogène entre deux variables quantitatives, entre une variable quantitative et

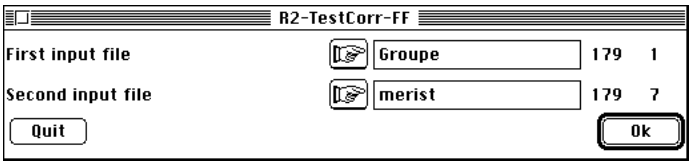
une variable qualitative, entre deux variables qualitatives. Trois utilitaires du module StatUtil permet de le faire.

2.1 — Lien quantitatif - quantitatif

Il est mesuré par la corrélation entre les deux variables et plus généralement par le carré de cette corrélation qui est le carré du cosinus des deux variables ou encore le pourcentage de variance expliquée par la régression de l'une sur l'autre. Il est testé classiquement par un test t [1].



Utiliser StatUtil : R2-TestCorr-FF :




```
Note : Two sides test t for correlation
Proba (X>|Obs| or X<-|Obs|)x10000 - R2x1000
i = number of variable from file merist
j = number of variable from file Groupe

[i =   1][j =   1] R =    0.08999 | Proba =  2288 | R2 =    8
[i =   1][j =   2] R =      0.234 | Proba =    18 | R2 =   55
[i =   1][j =   3] R =    -0.1026 | Proba =  1678 | R2 =   11
[i =   1][j =   4] R =    0.03243 | Proba =  6702 | R2 =    1
[i =   1][j =   5] R =  -0.009748 | Proba =  8925 | R2 =    0
[i =   1][j =   6] R =    0.07945 | Proba =  2907 | R2 =    6
[i =   1][j =   7] R =    -0.5717 | Proba =     0 | R2 =  327
```

On a ici utilisé le nombre d'allèles comme une variable quantitative.

2.2 — Lien qualitatif - quantitatif

Il est mesuré par la rapport de corrélation entre les deux variables qui est le carré du cosinus de la variable quantitative avec sa projection sur le sous-espace engendré par les indicatrices des classes, ou encore le pourcentage de variance expliquée par la courbe de régression [2]. Il est testé classiquement par une analyse de variance à un facteur contrôlé [3].

Utiliser StatUtil : R2-Anova1-FF :


```
Note : Test F for Anova one layout - Proba (F>Obs)x10000 - R2x1000
For details, use the option in Discrimin module
i = number of variable from file Groupe
j = number of variable from file merist

[i =   1][j =   1] F =      1.676 | Proba =  1286 | R2 =   55
[i =   1][j =   2] F =      2.248 | Proba =   407 | R2 =   73
[i =   1][j =   3] F =      1.262 | Proba =  2766 | R2 =   42
[i =   1][j =   4] F =     0.1824 | Proba =  9798 | R2 =    6
[i =   1][j =   5] F =     0.6705 | Proba =  6756 | R2 =   23
[i =   1][j =   6] F =     0.6483 | Proba =  6935 | R2 =   22
[i =   1][j =   7] F =      14.63 | Proba =     0 | R2 =  338
```

La variable 2, qui pourrait être intéressante, est en fait mal conditionnée et ne prend que trois valeurs distinctes (8, 9 et 10) dont deux sont rares. Sa signification est douteuse.

On a ici utilisé le nombre d'allèles comme une variable qualitative. Il n'y a pas de contradiction entre les deux séries de résultats mais il pourrait y en avoir en cas de liaisons non linéaires. La quasi-identité des R^2, carré de corrélation dans le premier cas (variable 7 : 0.327), rapport de corrélation dans le second (variable 7 : 0.338) indique une liaison linéaire stricte et confirme que la variable génétique est bien ici une variable quantitative.

Du bloc de variables méristiques, on extrait la dernière (CAEC Nombre de caeca pyloriques) dans le fichier M1 (197-1). On confirme un lien légèrement négatif avec la taille :


```
[i =   1][j =   1] R =   -0.1717 | Proba =  204 | R2 =    29
```

Ces deux approches des liens entre deux variables par le test de la corrélation ou le test de l'analyse de variance se résumant à un niveau de signification et un carré de cosinus (R^2) sont tout-à-fait élémentaire. Il est connu depuis longtemps qu'elles ont des équivalents stricts pour le cas de deux variables qualitatives. On n'en trouve que peu de traces dans la littérature francophone parce que cela implique de reconnaître que l'analyse des correspondances était déjà décrite clairement dans l'ouvrage classique de Kendall et Stuart [4], avec un test de signification.

	LTLog		G (qual)		G (quan)				LTLog		G (qual)		G (quan)	
CAEC	29	1	338	4	327	4		TPR	484	4	78	1	11	
								TPN	355	4	91	1	7	
PRAD	57	2	125	3	77	3		LS	986	4	115	2	39	2
PNAD	165	4	460	4	734	4		MD	968	4	126	3	55	2
PRAA	4		30		4			MAD	985	4	119	2	42	2
PNAA	175	4	312	4	268	4		MAN	980	4	110	2	36	1
PRDA	12		39		0			MPEL	957	4	105	2	28	1
PNDA	200	4	392	4	376	4		MPEC	871	4	122	2	58	2
PRLI	3		48		18			DAD	941	4	108	2	27	1
PRINF	49	2	54		8			DC	963	4	110	2	28	1
PNINF	76	3	527	4	500	4		DAN	951	4	111	2	22	1
PRSUP	0		45		15			DPEL	868	4	99	2	7	
PNSUP	212	4	341	4	321	4		DPEC	904	4	100	2	21	1
PNO	139	4	660	4	630	4		ADC	917	4	104	2	23	1
PRD	0		174	4	105	4		ADAN	669	4	117	2	20	
PND	243	4	205	4	191	4		ADPEL	936	4	125	3	35	1
								ADPEC	952	4	102	2	26	1
TâPéri	111	3	699	4	655	4		PECPE	880	4	78	1	7	
OcPR	177	4	76		36			PECAN	940	4	95	2	20	
OcPN	271	4	86		70	2		PECC	612	4	62		13	
MaJu	194	4	152	3	95	3		PELAN	899	4	99	2	36	1
TâOp	65	3	225	4	201	4		PELC	921	4	130	3	49	2
PNTêt	2		125	2	40	2		ANC	929	4	145	3	50	2
FrD	58	2	619	4	584	4		LPRO	861	4	128	3	43	2
PtsAd	53	2	33		17			DO	621	4	264	4	212	4
FrAd	3		30		2			LPOO	917	4	130	3	65	3
FrAn	11		408	4	367	4		LTET	924	4	151	1	91	3
FrPel	22		237	4	217	4		HTET	915	4	112	2	27	1
FrC	4		194	4	139	4		LMAX	851	4	170	4	116	4
PtsDos	3		45		13			LAD	755	4	153	1	97	4
CouFl	121	4	488	4	464	4		LD	877	4	112	2	39	2
ConPN	121	4	337	4	317	4		HD	619	4	117	2	82	3
Zéb	29	1	82		60	2		LC	753	4	123	3	46	2
								LAN	876	4	183	4	104	4
								HAN	827	4	176	4	120	4
								LPELG	831	4	183	4	126	4
								LPECG	716	4	191	4	158	4
								HPED	862	4	112	2	15	
								ETET	626	4	128	3	24	

Tableau 1 : Bilan par groupes de variables du lien avec la taille (LTLog) et la variable génétique G prise comme qualitative (qual) ou quantitative (quan). Le premier indicateur est un R2 (exprimé en millièmes) et le second est un niveau de signification statistique (rien : NS, 1 : <5%, 2 : <1%, 3 : <0.1%, 4 : <0.01%). Suivant le type de variables on utilise un test sur le coefficient de corrélation, l'analyse de variance ou le test de Kendall.

2.3 — Lien qualitatif - qualitatif

Quand deux variables sont qualitatives, on construit la table de contingence qui dénombre les individus par couple de modalités et on teste le résultat par un Khi2 à (*I*-1)(*J*-1) degrés de liberté si *I* et *J* sont les effectifs des modalités des deux variables [5]. Le test Khi2, généraliste, ne préjuge pas de l'origine de la signification et, en particulier est sensible dès qu'il y a une concentration d'individus dans une case du tableau. Ce qui est

plus intéressant est un test de la structure du tableau sur la première valeur propre de son AFC. La première valeur propre multipliée par le nombre total d'individus se teste par un Khi2 à *I*+*J*-3 degrés de liberté [6]. Comme cette première valeur propre est le cosinus carré de l'angle formé par les deux sous-espace engendré par les indicatrices, il a exactement le statut d'un carré de corrélation (on dit carré de corrélation canonique) ou d'un rapport de corrélation [7].

On caractérisera donc l'association entre deux variables qualitatives par le R2 canonique qu'on testera par le test de Kendall qu'on peut toujours comparer au Khi2. Ceci est disponible dans StatUtil : R2-Kendall-FF :


```
i = number of variable from file Groupe.cat
j = number of variable from file Photo.cat

Proba1 : Test Khi2 - Proba (Khi2>Obs)x10000
For details, use the option in CategVar module
Proba2 : Test Khi2 for canonical correlation
DF = I + J - 3
Cf. Kendall, D.G. & Stuart, A. (1961)
The advanced theory of statistics.
Vol 2: Inference and relationships.
Cha. 33 : Categorized data. Griffin, London. 536-591.
R2x1000 : Canonical correlation = first eigenvalue * tot

[i =   1][j =   1] Proba1 =     0 | Proba2 =     0 | R2 =  699
[i =   1][j =   2] Proba1 =   962 | Proba2 =   591 | R2 =   76
[i =   1][j =   3] Proba1 =   811 | Proba2 =   304 | R2 =   86
[i =   1][j =   4] Proba1 =     2 | Proba2 =     2 | R2 =  152
[i =   1][j =   5] Proba1 =     0 | Proba2 =     0 | R2 =  225
[i =   1][j =   6] Proba1 =    12 | Proba2 =    12 | R2 =  125
[i =   1][j =   7] Proba1 =     0 | Proba2 =     0 | R2 =  619
[i =   1][j =   8] Proba1 =  4281 | Proba2 =  4281 | R2 =   33
[i =   1][j =   9] Proba1 =  4902 | Proba2 =  4902 | R2 =   30
[i =   1][j =  10] Proba1 =     0 | Proba2 =     0 | R2 =  408
[i =   1][j =  11] Proba1 =     0 | Proba2 =     0 | R2 =  237
[i =   1][j =  12] Proba1 =     0 | Proba2 =     0 | R2 =  194
[i =   1][j =  13] Proba1 =  2361 | Proba2 =  2361 | R2 =   45
[i =   1][j =  14] Proba1 =     0 | Proba2 =     0 | R2 =  488
[i =   1][j =  15] Proba1 =     0 | Proba2 =     0 | R2 =  337
[i =   1][j =  16] Proba1 =   233 | Proba2 =   233 | R2 =   82
```

Quand une des variables a deux modalités seulement les deux tests sont identiques. Sinon ils sont ici cohérents. On peut comparer avec l'approche de la variable génétique comme variable quantitative :


```
Note : Test F for Anova one layout - Proba (F>Obs)x10000 - R2x1000
For details, use the option in Discrimin module
i = number of variable from file Photo
j = number of variable from file Groupe

[i =   1][j =   1] F =      166.9 | Proba =     0 | R2 =  655
[i =   2][j =   1] F =      3.263 | Proba =   396 | R2 =   36
[i =   3][j =   1] F =      6.622 | Proba =    19 | R2 =   70
[i =   4][j =   1] F =      18.55 | Proba =     1 | R2 =   95
[i =   5][j =   1] F =      44.53 | Proba =     0 | R2 =  201
[i =   6][j =   1] F =      7.374 | Proba =    72 | R2 =   40
```

```
[i =   7][j =   1] F =      248.1 | Proba =     0 | R2 =  584
[i =   8][j =   1] F =      2.985 | Proba =   819 | R2 =   17
[i =   9][j =   1] F =     0.3445 | Proba =  5654 | R2 =    2
[i =  10][j =   1] F =      102.5 | Proba =     0 | R2 =  367
[i =  11][j =   1] F =      24.46 | Proba =     0 | R2 =  217
[i =  12][j =   1] F =      28.64 | Proba =     0 | R2 =  139
[i =  13][j =   1] F =      2.243 | Proba =  1318 | R2 =   13
[i =  14][j =   1] F =      153.5 | Proba =     0 | R2 =  464
[i =  15][j =   1] F =      82.04 | Proba =     0 | R2 =  317
[i =  16][j =   1] F =      11.28 | Proba =    11 | R2 =   60
```

La liste des variables hautement significatives est la même, ce qui confirme encore qu'on a moins une partition qu'un gradient génétique. L'identité entre les deux séries est remarquable.

On utilise les trois tests pour établir le tableau 1. L'omniprésence des deux variables de contrôle taille et groupe génétique et la cohérence totale entre variable génétique vue comme partition et variable génétique quantitative sont les principaux résultats. Cela ne simplifie pas le problème. On se retrouve avec deux variables instrumentales quantitatives : la première est parasite (taille) et la seconde est au centre de l'étude (génétique). On veut éliminer l'effet de la première et mettre en évidence l'effet de la seconde, avec deux contraintes. La première est qu'elles sont liées et la seconde est que cet effet porte sur un ensemble de descripteurs de plusieurs types (quantitatif ou qualitatif) liés entre eux dans un type donné (voir les analyses préliminaires) et liées entre types (Cf. ci-après). On va d'abord montrer que les types de variables peuvent se mélanger sans difficulté.

3 — Mélanges d'ACP et d'ACM

La question porte sur le mélange des types de variables. Supposons qu'un ensemble de descripteurs forment un groupe de signification biologique comme la description de la robe de l'animal. Ces descripteurs peuvent être quantitatifs comme le dénombrement des points rouges et noirs (fichier ponctua) ou qualitatifs (présence ou absence de tels ornements, ..., fichier Photo).

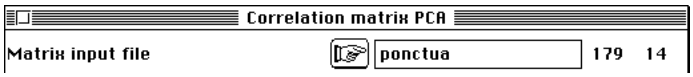
Il est clair que c'est la seule contrainte technique qui sépare les deux groupes. Il est par contre légitime de séparer les descripteurs de la robe de ceux de la forme de l'animal (variables morphométriques) mais c'est encore une contrainte technique qui sépare les descripteurs méristiques indépendants de la taille et les paramètres dimensionnels extrêmement dépendants de la taille du poisson.

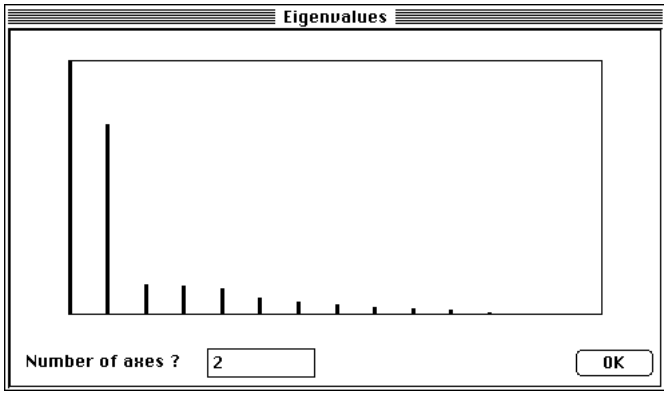
On voudrait ranger les descripteurs par type sur des critères biologiques et non sur des critères statistiques. De manière plus générale, la notion de variable vue par la statistique (une colonne d'un tableau) et la notion de variables vue par la biologie (un type de renseignement contenu éventuellement sur plusieurs colonnes d'un tableau) sont souvent divergents. Le codage flou [8] de l'information est l'archétype de cette difficulté. Il faudra, à terme, utiliser des méthodes multivariées qui tolère à égalité les variables quantitatives d'unité quelconque (variables susceptibles de la normalisation et de l'ACP sur matrices de corrélation), les variables qualitatives strictes (relevant de l'ACM classique) et les variables floues (codant une association individu-modalités non stricte, c'est-à-dire un preferendum comme le profil écologique). La notion à égalité veut dire qu'une variable quantitative qui compte pour une colonne doit être de même poids qu'une variable qualitative ou floue quelque soit le nombre des modalités.

Des solutions techniques (analyses canoniques généralisées) existent dans la littérature statistique sans être connues en biologie. Pour y introduire nous décrirons d'abord l'analyse de Hill & Smith (1976) [9].

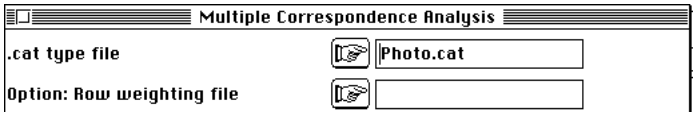
On utilise l'option MCA : Hill & Smith Analysis pour mixer une ACP et une ACM.

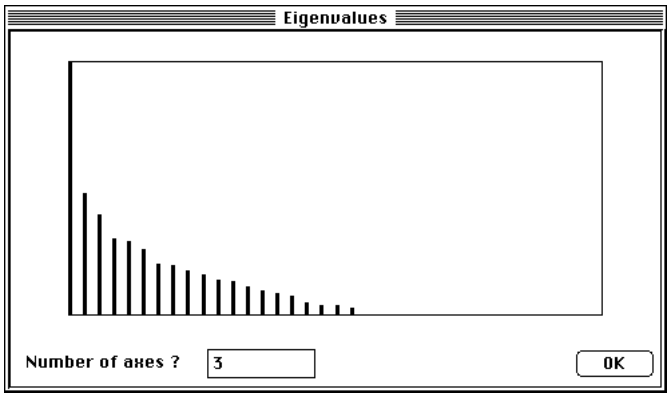
Commencer par l'exécution de l'ACP normée du tableaux de variables quantitatives :



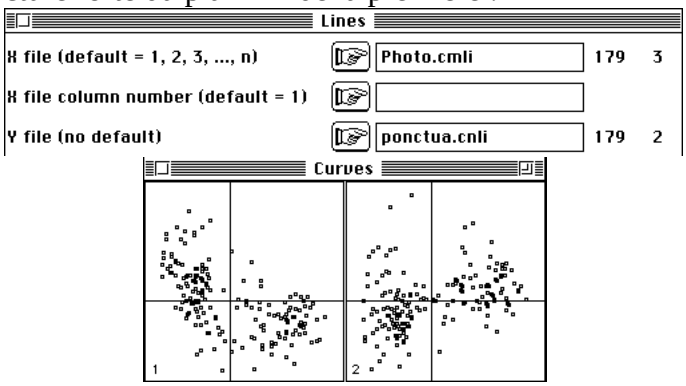


Continuer par l'exécution de l'ACM du tableau de variables qualitatives :

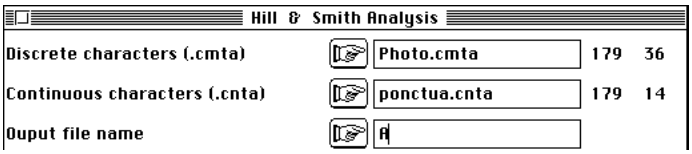


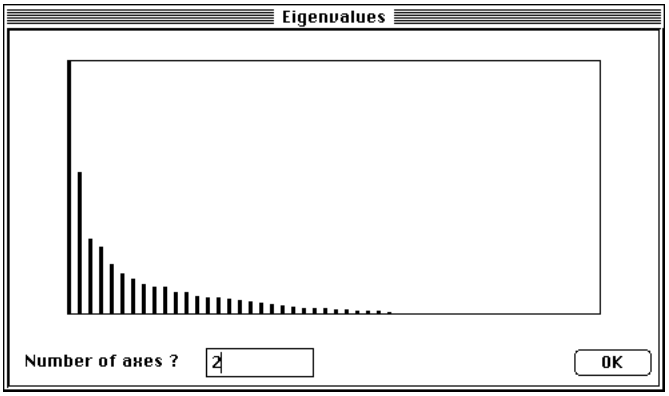


Il serait manifestement intéressant si la structure forte du facteur 1 de la seconde se retrouve sur la structure forte du plan 1-2 de la première :



Pour le mettre en évidence, il suffit de mélanger les deux analyse :






```
cm/MCA: Hill & Smith analysis
Hill, M.O. & Smith, A.J.E. (1976)
Principal component analysis of taxonomic data with multi-state discrete
characters
Taxon : 25, 249-255
```

L'option exécute strictement les propositions théoriques des auteurs.

```
First Input (Multiple correspondence analysis): Photo.cmta
Second Input (Normed principal component analysis): ponctua.cnta
Output table (Mixed): A.hita
File A.hita has 179 rows and 50 columns (36 categories + 14 variables)
```

On constitue un nouveau triplet statistique en accolant purement et simplement les tableaux des deux analyses de base. Le programme contrôle que la compatibilité est assurée en vérifiant que les pondérations des analyses de départ sont identiques. Cette pondération commune des lignes des tableaux de départ est conservée comme pondération des lignes du nouveau tableau :

```
File A.hipl contains the row weights
It has 179 rows and 1 column
```

Les pondération des colonnes sont par contre modifiées. Dans l'ACM chaque colonne est une modalité. On somme les poids des porteurs de cette modalité et on divise par le nombre de variables. Ceci attribue des poids aux modalités tels que chaque variable totalise (par ses modalités) un poids égal à un sur le nombre de variables. On fait ici la même chose mais en divisant par le nombre de variables total (qualitatives et quantitatives) et au attribue aux variables quantitatives le poids uniforme de un sur le nombre total de variables. Ici, il y a 16 variables qualitatives et 14 variables quantitatives. Chacune des dernières a un poids de 1/30 et les poids des modalités des premières sommés sur par variable redonne 1/30. La somme totale des poids des colonnes du tableau vaut ainsi 1.

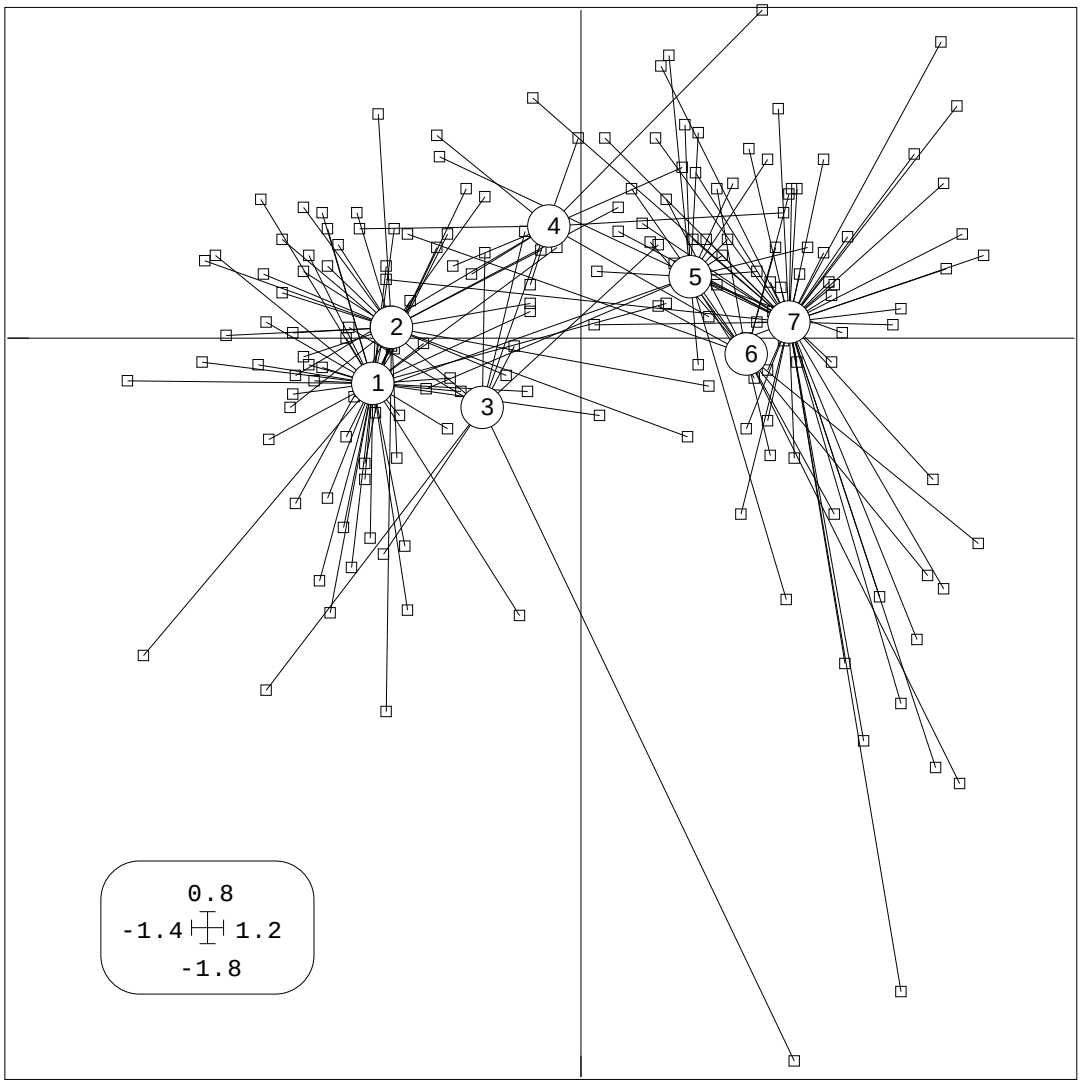
```
File A.hipc contains the column weights (V/nvartot)*DM and (1/nvartot)Ip
It has 50 rows  (36 categories + 14 variables) and 1 column
```

On obtient ainsi un nouveau triplet statistique qui donne une analyse d'inertie standard. L'inertie de l'ACM initiale (nombre de modalités/nombre de variables -1) vaut 36/16 -1, soit 1.25. L'inertie de l'ACP initiale vaut (nombre de variables) 14. L'inertie de l'analyse conjointe vaut 1.25(16/30) +14/30 = 1.1333. Comme en ACP normée ou en ACM cette valeur ne dépend pas des observations mais de la structure des variables.

```
DiagoRC: General program for two diagonal inner product analysis
Input file: A.hita
--- Number of rows: 179, columns: 50
----------------------
Total inertia:  1.13333
----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +2.9891E-01 +0.2637 +0.2637   |02   +1.6694E-01 +0.1473 +0.4110 |
03   +8.8128E-02 +0.0778 +0.4888   |04   +7.9174E-02 +0.0699 +0.5587 |
05   +5.9643E-02 +0.0526 +0.6113   |06   +4.8061E-02 +0.0424 +0.6537 |
•••
33   +4.8337E-04 +0.0004 +0.9998   |34   +2.0357E-04 +0.0002 +1.0000 |
35   +0.0000E+00 +0.0000 +1.0000   |36   +0.0000E+00 +0.0000 +1.0000 |
37   +0.0000E+00 +0.0000 +1.0000   |38   +0.0000E+00 +0.0000 +1.0000 |
39   +0.0000E+00 +0.0000 +1.0000   |40   +0.0000E+00 +0.0000 +1.0000 |

File A.hivp contains the eigenvalues and relative inertia for each axis
--- It has 50 rows and 2 columns
```

On a gardé deux axes au vu du graphe des valeurs propres parfaitement expressif. On obtient des coordonnées des lignes et des coordonnées des colonnes.

Les coordonnées des lignes sont centrées, de variance égales aux valeurs propres. On s'en sert comme dans une ACM ou une ACP :



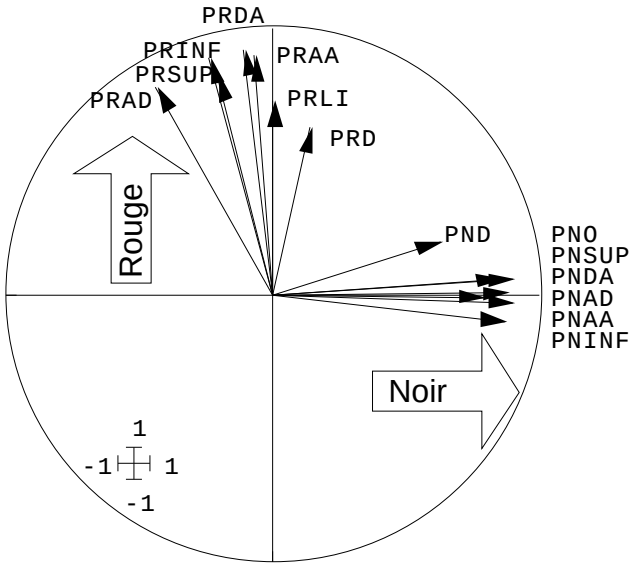

```
File A.hili contains the row scores
--- It has 179 rows and 2 columns
```

```
File :A.hili
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.106e+00| 9.854e-01|
|   2|-1.761e+00| 7.999e-01|
|----|----------|----------|
```

L'axe 1 est sensiblement un axe de discrimination génétique avec un pattern de continuum habituel. L'axe 2 est orthogonal et indépendant de la génétique. Les coordonnées des colonnes ont été décomposée en deux parties pour les manipulations d'une part des modalités des qualitatives, d'autre part des quantitatives. Toute sorte d'approches graphiques sont ainsi possibles. Les colorations rouge et noire sont à nouveau isolées et recalées sur le pattern des variables qualitatives.

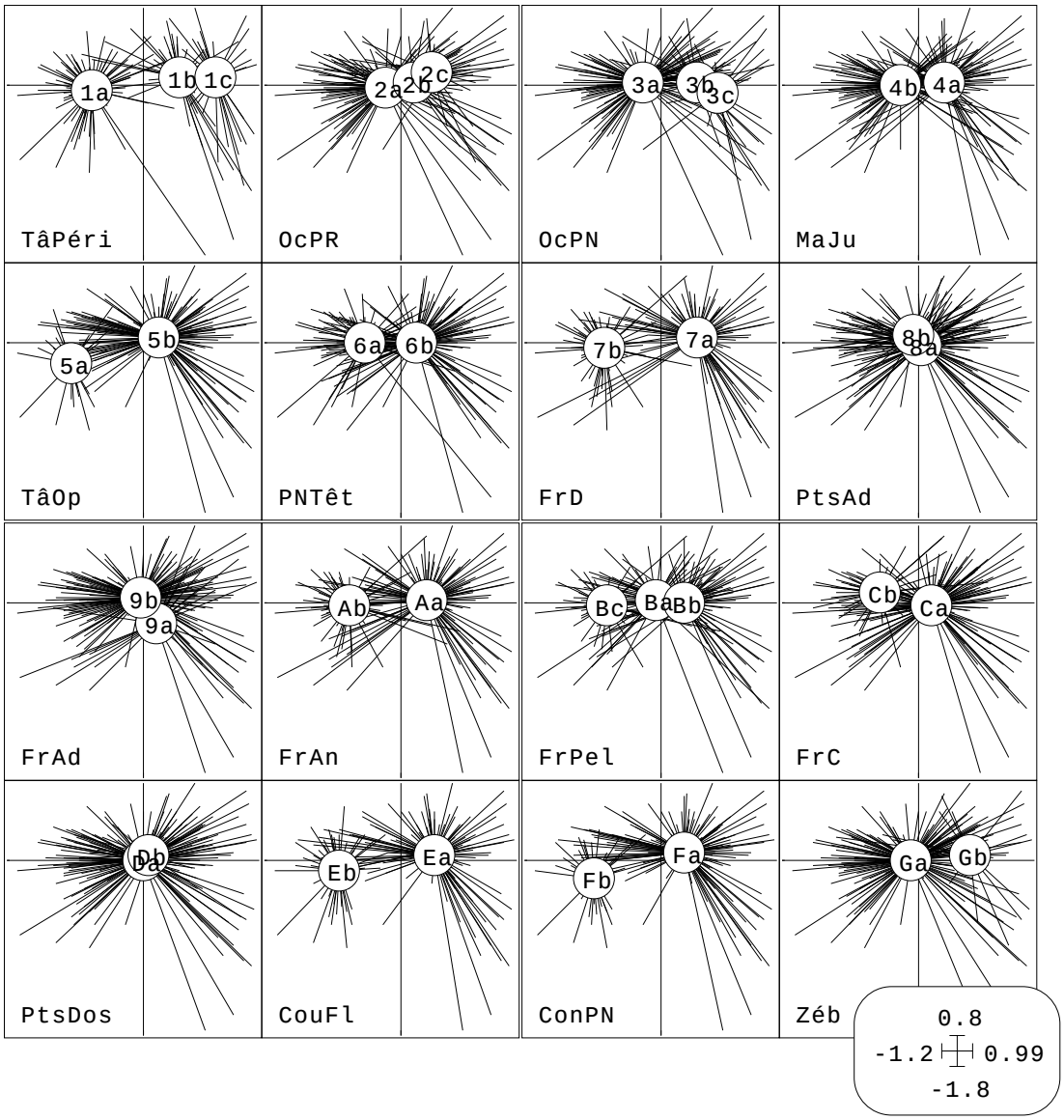
```
File A.hico contains the column scores
--- It has 50 rows and 2 columns
File :A.hico
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.134e+00| 1.140e+00|
|   2|-5.477e-01| 9.176e-01|
|----|----------|----------|

File A.hicocont contains the column scores of continuous parameters
--- It has 14 rows and 2 columns
--- It contains 14 last lines from file A.hico
--- It can be used for drawing correlation circles
File :A.hicocont
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-4.378e-01| 9.026e-01|
|   2|-1.067e-01| 9.176e-01|
|----|----------|----------|
```

L'interprétation pour les variables quantitatives se fait comme en ACP :




```
File A.hicocate contains the column scores of categories (discrete
parameters)
--- It has 36 rows and 2 columns
--- It contains 36 first lines from file A.hico
--- It is to be used with 43316613 and associated files
File :A.hicocate
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.134e+00| 1.140e+00|
|   2|-5.477e-01| 2.778e-01|
|----|----------|----------|
```

L’interprétation pour les variables qualitatives se fait comme en ACM :



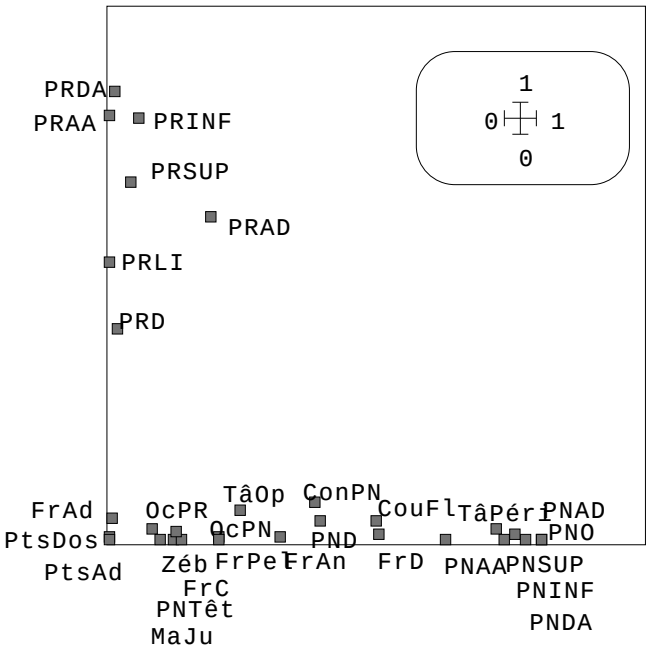
On remarque que l’axe 1 de l’ACM est venu se caler sur la coloration noire et qu’il n’y a pas d’équivalent dans les variables photographiques de la coloration rouge, ce qui est un résultat particulièrement net. Enfin, la théorie des cosinus carrés des angles donne à cette analyse des propriétés numériques extrêmement efficaces. La coordonnée des lignes (axe 1) en ACP normée maximise la somme des carrés des corrélations entre les variables et un score (propriété rappelée dans [9] et illustrée dans [10]). La coordonnée des lignes (axe 1) en ACM maximise la moyenne des rapport de corrélation entre les variables qualitatives et un score (propriété due à [11], exprimée en terme d’analyse de variance dans 9 et illustrée dans [12]). La présente analyse réunit les deux propriétés et maximise la moyenne sur l’ensemble des variables des R^2 entre variables et un score, ce R^2 étant un carré de corrélation pour les quantitatives et un rapport de corrélation pour les qualitatives. Cette moyenne sur l’ensemble des descripteurs vaut la valeur propre.

Le listing donne ces quantités pour chaque axe et on y retrouve la relation des deux ensembles de variables sur l’axe 1 et la seule prise en compte d’un sous-ensemble de quantitatives sur l’axe 2, ce qui vient d’être exprimée sur les graphiques précédents.

```
R2 (x1000) Column = axes
First bloc: discrete parameters
```

```
Second bloc: continuous parameters
Third bloc: overall mean = eigenvalue

Variable :   1       729      24
Variable :   2        84      20
Variable :   3       210       7
Variable :   4       124       3
Variable :   5       248      58
Variable :   6       138       0
Variable :   7       507      14
Variable :   8         3       9
Variable :   9         6      44
Variable :  10       324       5
Variable :  11       206       4
Variable :  12       128      16
Variable :  13         1       3
Variable :  14       503      37
Variable :  15       385      72
Variable :  16        97       2

Variable :   1       192     607
Variable :   2       814       1
Variable :   3         4     800
Variable :   4       633       0
Variable :   5        11     842
Variable :   6       784       0
Variable :   7         0     522
Variable :   8        56     790
Variable :   9       764      11
Variable :  10        42     674
Variable :  11       740       3
Variable :  12       815       4
Variable :  13        20     396
Variable :  14       398      39

Overall mean         299     167

File A.hiR2 contains the R2 coefficients
--- It has 30 rows and 2 columns
```

La lisibilité des résultats dans l'analyse de Hill et Smith est donc aussi bonne que dans une ACP normée simple et que dans une ACM simple. Cette analyse est donc parfaitement adaptée aux mélanges de types de descripteurs. Le dernier fichier permet de représenter les caractères quelle que soit leur type :



Toutes les variables associées à la coloration rouge peuvent désormais être exclues de la discussion. On notera cependant un fait marquant : la coloration noire est fortement liée à la taille du poison alors que la coloration rouge globalement n'en dépend pas. De même la coloration noire est fortement associée à la variable génétique alors que la coloration rouge en est totalement indépendante. Dans le bloc de variables sur la ponctuation, il y a deux groupes de paramètres de signification biologique totalement distinctes. Un même type technique de variables peut recouvrir plusieurs types biologiques alors que plusieurs types techniques peuvent former un ensemble cohérent quant à sa signification.

La coloration noire est fortement liée aux descripteurs qualitatifs de la robe qui sont eux-mêmes fortement liés à la taille du poisson. L'omniprésence de l'effet taille impose donc un effort sérieux pour s'en débarrasser. Pour continuer cette étude nous utiliserons l'analyse canonique généralisée dont l'analyse de Hill et Smith est un cas particulier.

Références

[1] Dagnelie, P. (1975) *Théories et méthodes statistiques.* Volume II. *Les méthodes de l'inférence statistique*. Les Presses Agronomiques de Gembloux, Gembloux. 1-362. p. 311.

[2] Rouanet, H. & Le Roux, B. (1993) *Analyse des données multidimensionnelles*. Dunod, paris. 1-310. p. 95-96.

[3] Dagnelie, P. (1975) p. 117.

[4] Kendall, D.G. & Stuart, A. (1961) *The advanced theory of statistics*. Vol 2: *Inference and relationships*. Cha. 33 : *Categorized data*. Griffin, London. 536-591.

[5] Dagnelie, P. (1975) p. 82.

[6] Kendall, D.G. & Stuart, A. (1961) p. 574.

[7] Lebart, L., Morineau, A. & Piron, M. (1995) Statistique exploratoire multidimensionnelle. Dunod, Paris. 1-439. p. 219.

[8] Chevenet, F., Dolédec, S. & Chessel, D. (1994) A fuzzy coding approach for the analysis of long-term ecological data. *Freshwater Biology* : 31, 295-309.

[9] Hill, M.O. & Smith, A.J.E. (1976) Principal component analysis of taxonomic data with multi-state discrete characters. *Taxon* : 25, 249-255.

[10] Carrel, G., Barthelemy, D., Auda, Y. & Chessel, D. (1986) Approche graphique de l'analyse en composantes principales normée : utilisation en hydrobiologie. *Acta Œcologica, Œcologia Generalis* : 7, 2, 189-203.

[11] Saporta, G. (1975) *Liaisons entre plusieurs ensembles de variables et codage de données qualitatives*. Thèse de 3° cycle, Université Pierre et Marie Curie, Paris VI. 1-102.

[12] Pialot, D., Chessel, D. & Auda, Y. (1984) Description de milieu et analyse factorielle des correspondances multiples. *Compte rendu hebdomadaire des séances de l'Académie des sciences*. Paris, D : 298, Série III, 11, 309-314.