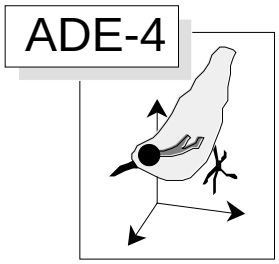


Analyses Discriminantes

Résumé

Parmi de nombreuses méthodes de discrimination, l'analyse discriminante linéaire, ou analyse factorielle discriminante est la plus connue. Elle sert essentiellement à décrire ce qui distingue les moyennes par groupe de plusieurs variables mesurées sur plusieurs individus de plusieurs groupes. Si on ne poursuit pas cet objectif, la méthode est sans signification. La fiche décrit l'emploi du module Discrimin. En illustrant un problème de morphométrie avec l'analyse discriminante on remonte aux sources (Mahalanobis, P.C. (1936) On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India* : 12, 49-55). Le programme est cependant beaucoup plus général et on l'emploie après une ACP à centrage multiplicatif, ce qui est inhabituel.

Plan

S. Dolédec, D. Chessel et H. Persat

1 — Présentation du problème

La fiche décrit l'analyse d'un tableau de morphométrie[1][2]. 120 individus d'une espèce (l'Ombre commun, *Thymallus thymallus*) proviennent de cinq populations et portent 13 mesures. L'objectif est la description des différences entre populations, c'est-à-dire la discrimination inter-populationnelle. On insistera sur la présence de l'effet taille[3].

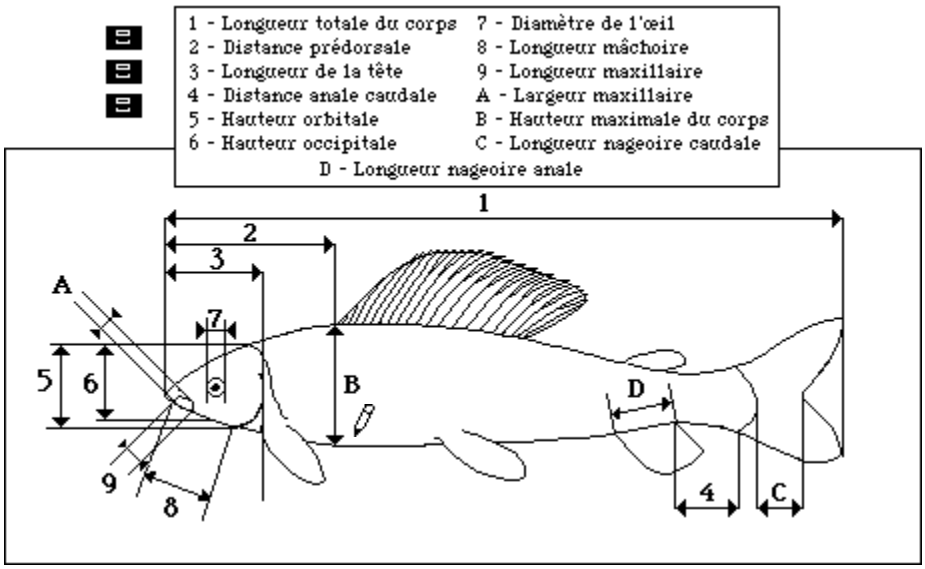


Figure 1 — Définition des 13 variables mesurées sur 120 individus.

Les variables mesurées sont décrites sur la figure 1. Les données traîtées sont dans le tableau 1. Les individus sont rangés par lieu de capture :

```
1- [  1- 41] Ain amont (41 individus)
2- [ 42- 59] Bienne (18 individus)
3- [ 60- 79] Loue (20 individus)
4- [ 80-102] Ain aval (23 individus)
5- [103-120] Loire (18 individus)
```

On note **Y** le tableau des données brutes, *n* (120) le nombre de lignes, *p* (13) le nombre de variable, **q** la variable qualitative qui indique pour chaque individu la classe à laquelle il appartient, *m* le nombre de classe, **Q** le tableau disjonctif complet associé à cette partition (*n* lignes et *m* colonnes définies par les indicatrices des classes), **D** la matrice diagonale (*n* lignes et *n* colonnes) contenant le poids des individus choisis *a priori*. Ici aucune raison ne conteste le choix implicite d'une pondération uniforme (chaque ligne de **Y** a pour poids 1/*n*).

La question de la discrimination descriptive se résume à la recherche de ce qui permet de distinguer les individus des différentes populations, éventuellement en précisant si cette question est valide (présence de différences significatives). La question est largement compliquée par l'effet taille, une séparation des populations par la taille des individus étant considérée comme de peu d'intérêt.

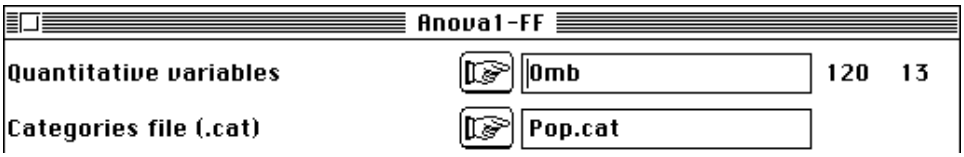
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	299	90	49	30	297	369	138	216	132	54	552	190	240
2	297	89	47	31	270	370	136	212	124	64	523	170	238
3	335	99	52	34	293	419	119	236	136	67	618	190	289
4	311	97	48	36	270	359	138	216	118	62	548	210	238
5	323	97	54	31	288	405	142	228	130	66	599	180	261
6	340	108	55	32	324	423	161	233	148	66	587	180	266
7	301	89	49	31	282	364	139	216	124	68	542	200	239
8	335	101	55	31	293	393	152	228	140	66	617	210	273
9	328	97	52	32	296	392	143	223	136	62	648	190	277
10	296	92	48	36	239	334	136	203	118	55	554	220	237
11	296	92	46	27	266	342	135	205	114	48	541	190	237
12	304	93	50	31	273	380	136	206	114	63	560	200	241
13	306	93	48	33	274	368	149	212	133	65	562	200	238
14	306	91	48	32	272	394	142	214	121	65	596	180	254
15	313	96	49	33	259	370	147	211	124	61	596	170	253
16	369	114	56	38	305	460	150	228	135	71	758	200	308
17	294	92	49	31	248	357	134	216	120	60	557	170	214
18	301	92	49	30	267	340	139	214	121	63	578	180	245
19	317	93	48	37	276	374	141	214	125	59	572	170	247
20	309	96	50	33	242	360	136	210	130	64	622	200	242
21	343	103	56	33	289	412	148	249	140	67	660	210	224
22	334	102	54	33	277	342	142	239	134	61	600	220	277
23	331	101	56	33	272	369	161	221	140	62	640	190	273
24	299	92	50	31	246	347	144	215	125	55	535	160	235
25	345	103	55	34	289	386	156	236	132	69	669	210	288
26	283	88	46	29	241	334	140	294	115	50	499	190	214
27	314	97	51	31	290	400	143	235	122	66	570	210	240
28	306	90	51	30	270	355	144	225	122	55	540	160	230
29	341	105	57	32	300	410	150	250	140	66	630	210	310
30	344	104	55	34	340	458	150	244	140	67	650	190	283
...													
103	250	71	41	23	224	327	113	190	103	48	447	130	208
104	240	68	41	24	213	310	115	187	101	47	447	140	223
105	239	68	39	24	206	280	115	176	101	40	423	120	192
106	245	74	41	24	212	290	120	185	99	45	422	130	216
107	247	73	41	24	211	284	115	186	100	44	425	140	207
108	249	69	40	25	227	322	114	184	104	46	431	130	196
109	319	90	53	28	282	416	139	235	134	62	616	180	294
110	299	85	48	29	245	368	125	218	117	55	577	160	248
111	247	72	41	24	203	291	115	183	105	45	450	140	201
112	245	70	40	23	200	307	115	179	100	44	450	140	198
113	313	88	49	30	256	357	133	222	118	66	564	160	277
114	231	70	40	22	213	303	123	171	98	42	422	130	200
115	269	72	42	27	227	332	123	187	102	48	496	140	249
116	245	71	41	23	218	303	117	177	95	44	439	120	200
117	262	74	44	26	222	324	121	191	108	47	478	140	247
118	264	77	44	29	229	326	122	190	104	52	482	150	244
119	274	76	43	28	231	337	124	193	106	51	503	140	251
120	243	69	41	23	211	301	117	180	97	49	435	130	209

Tableau 1 — 13 mesures enregistrées sur 120 poissons. Données des populations 1 et 5.

2 — Analyses univariées

Implanter les données de la carte Ombres. Le fichier Omb.txt donne un fichier binaire Omb (120 lignes et 13 colonnes) et le fichier Pop.Car donne un fichier Pop (120 lignes et 1 colonne). Lire ce fichier avec ReadCateg :

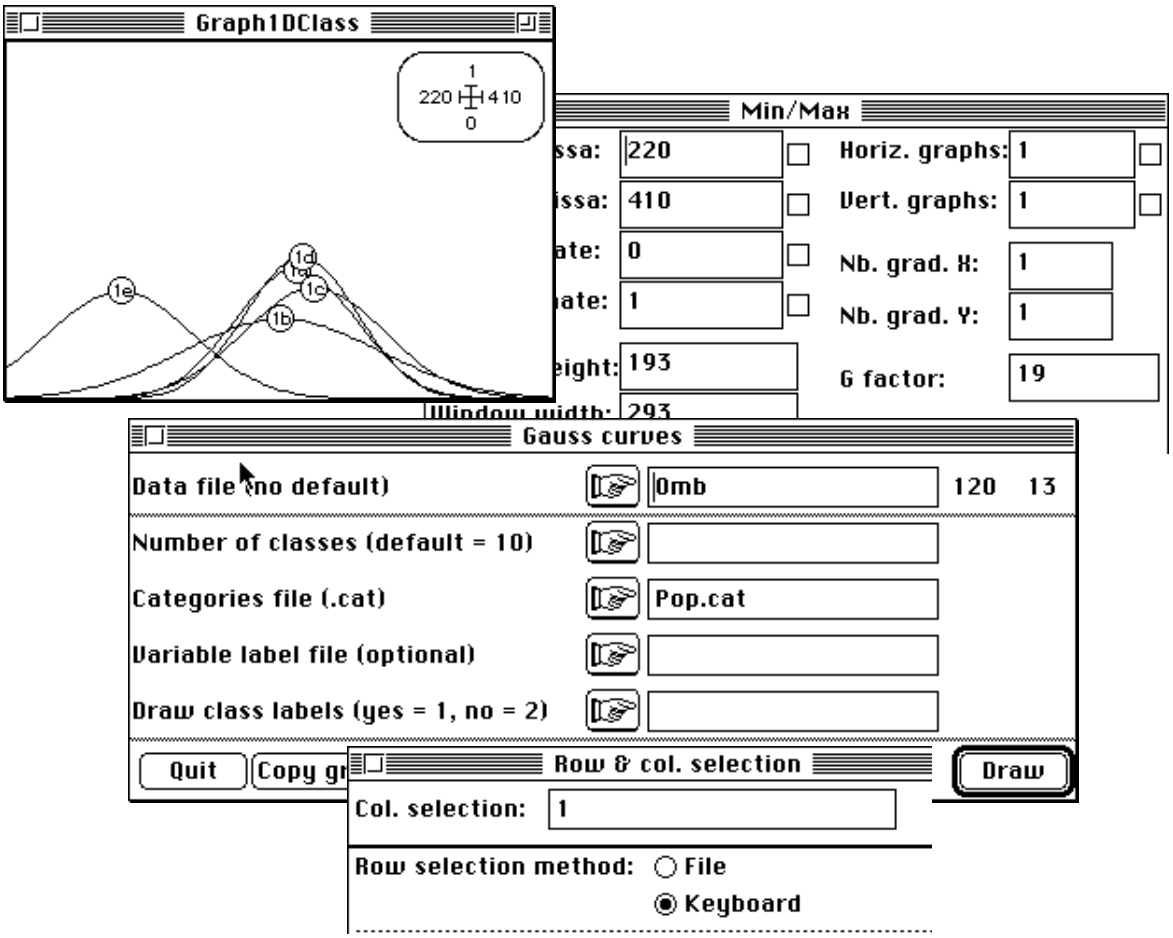
Utiliser dans Discrimin l’option :

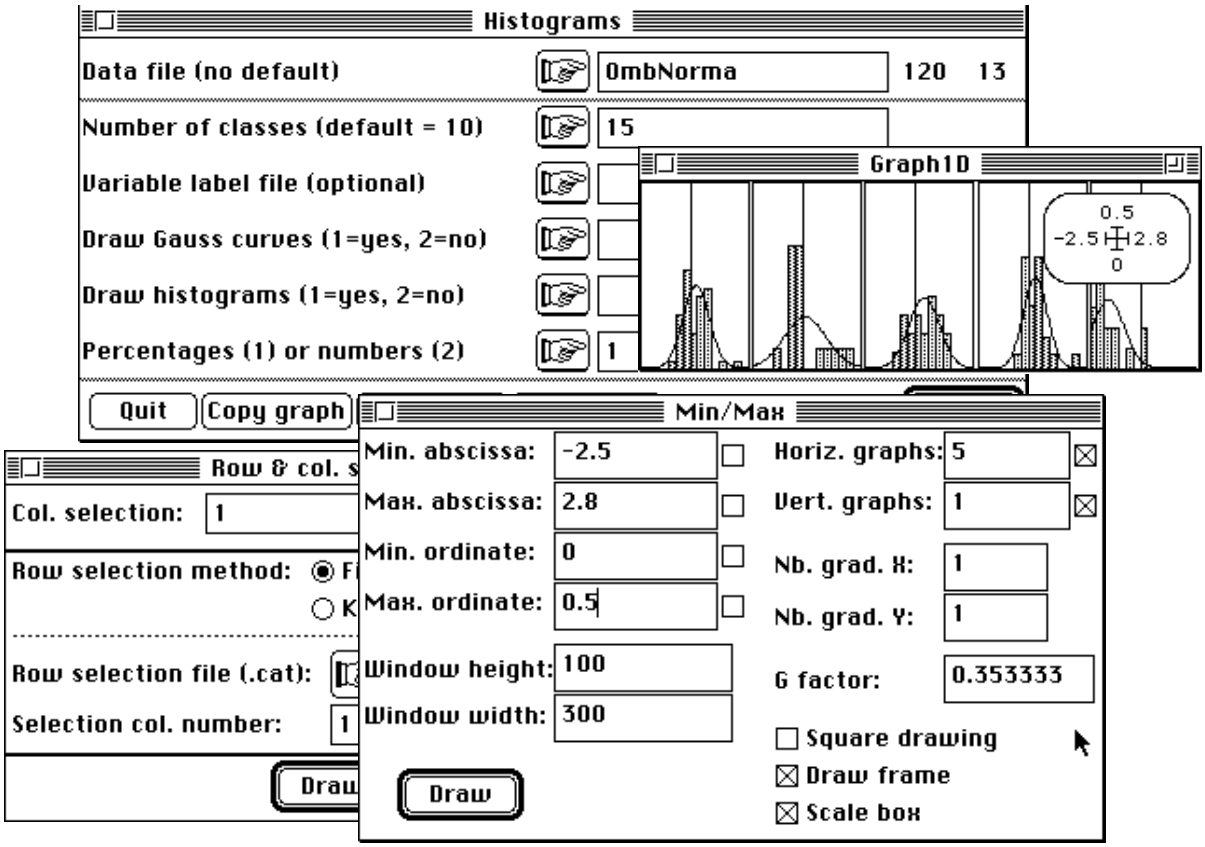


La première approche concerne la capacité de chacune des variables à discriminer les 5 groupes. On vérifie d'abord que certaines variables discriminent significativement les groupes. C'est le rôle de l'analyse de variance univariée :

```
variable 1 from Omb versus variable 1 from Pop
------------------------------------------------------------
|Source  |        SS|d.f.|        MS|         F|     Proba|
|Between | 5.946E+04|   4| 1.486E+04|     24.37|         0|
|Within  | 7.013E+04| 115|     609.8|
|Total   | 1.296E+05| 119|
------------------------------------------------------------

variable 2 from Omb versus variable 1 from Pop
------------------------------------------------------------
|Source  |        SS|d.f.|        MS|         F|     Proba|
|Between |      7918|   4|      1980|     37.92|         0|
|Within  |      6003| 115|      52.2|
|Total   | 1.392E+04| 119|
------------------------------------------------------------
•••
```





On y observe que chacune des variables permet de rejeter sans risque l'hypothèse d'égalité des moyennes dans les sous-populations. Une description multivariée est donc légitime.

Une vue descriptive de la relation entre la variable de tri (partition) et chacune des variables mesurées est possible avec des histogrammes pour chaque variable et chaque classe avec une échelle spécifique. Les variables 1 à 5 sont utilisées sur la figure 2.

Le résumé par des courbes de Gauss ne soulève pas d'objections et on peut superposer ces modèles (Figure 3). La population 5 (Loire) y apparaît systématiquement avec des individus plus petits sur toutes les variables. La plus grande partie de la signification des analyses de variance est attachée à cet élément.

Le fait que les individus d'une population soient systématiquement plus petits que ceux des autres, lorsque l'age des individus n'est pas contrôlé a un intérêt limité. En outre, la différence peut ne dépendre que des conditions trophiques, lesquelles ne sont pas non plus maîtrisées. Il est ici question de mesurer entre individus d'une population et entre individus de différentes populations des différences de forme et non de taille. Ces différences de forme donneront, à coup sûr, de meilleurs renseignements sur l'identité génétique des populations, leur homogénéité et leur origine, la présence de mélanges et le rôle des alevinages au coeur des problèmes posés par le traitement de ce tableau.

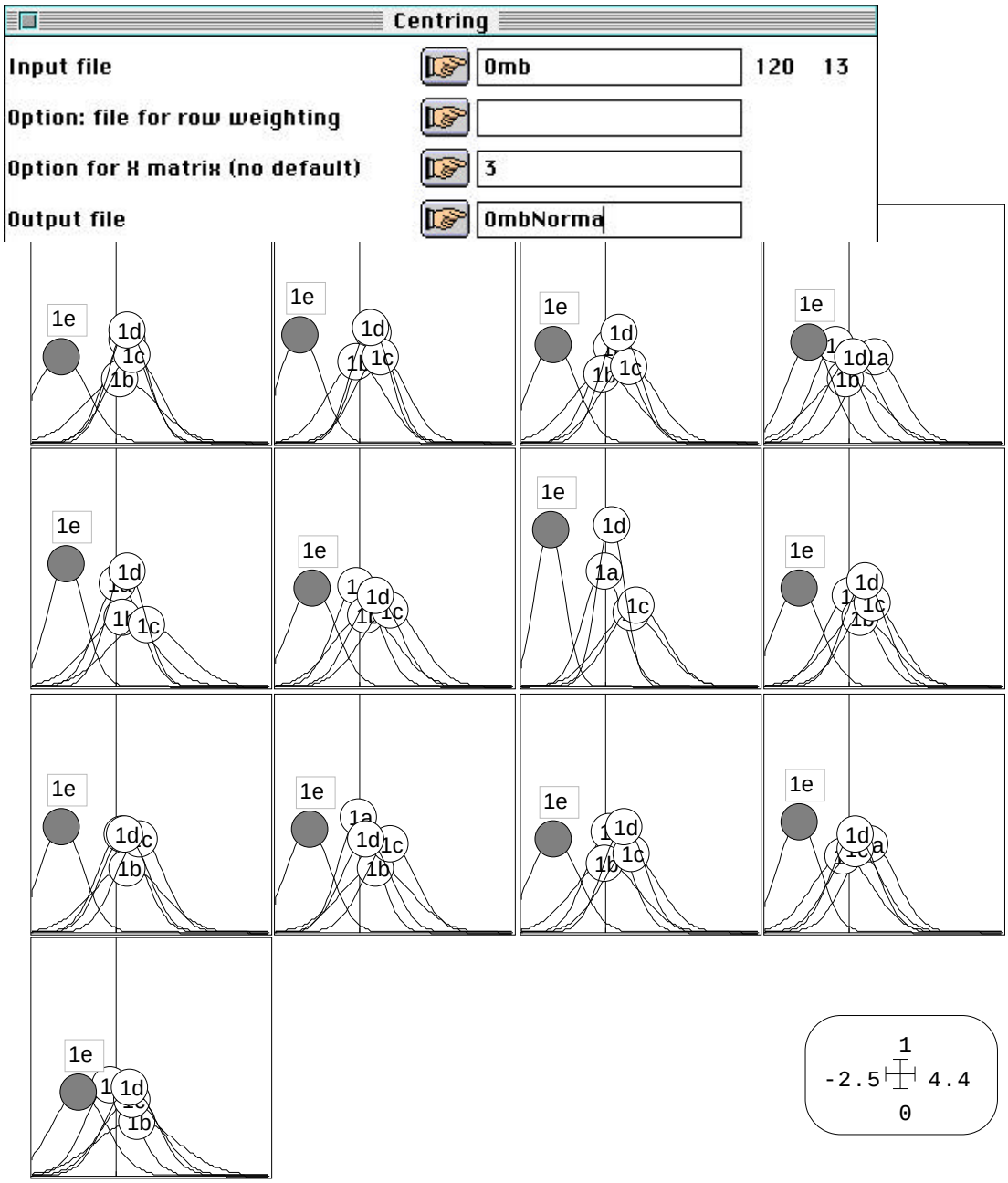


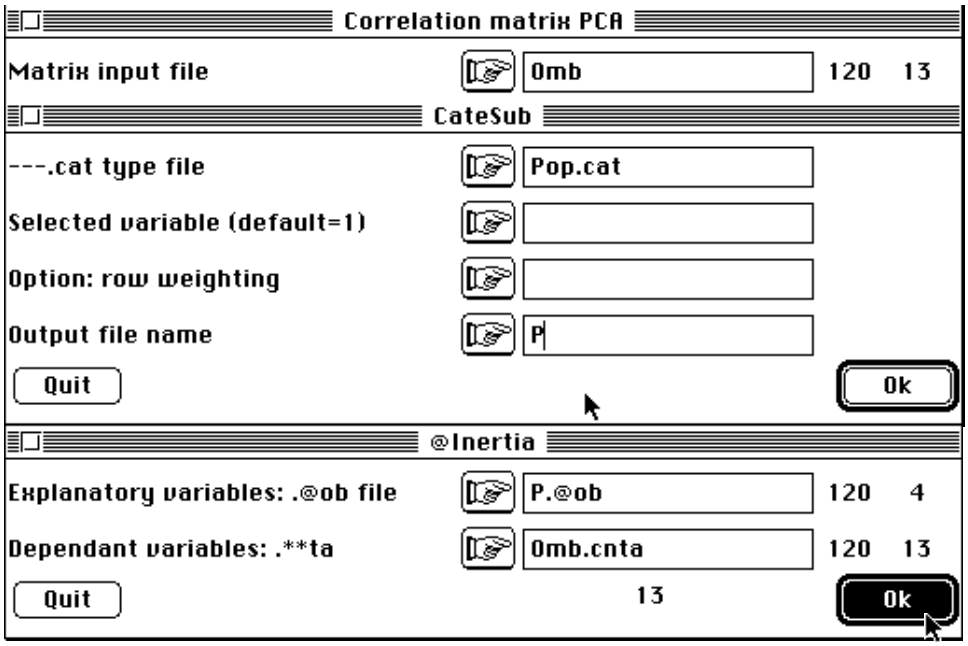
Figure 2 — Expression des moyennes et des variances de chaque groupe et de chaque variable (données normalisées par l'option Centring de Bin->Bin). Module graphique Graph1DClass.

La définition même de la forme est un problème complexe auquel introduit avec efficacité la revue récente de YOCCOZ [4].

Au plan descriptif, on peut résumer la participation de chaque variable à la séparation des populations par le pourcentage de variance inter-classes ou rapport de corrélation :

```
| 1 : 0.4588  | 2 : 0.5688  | 3 : 0.4534  | 4 : 0.4024 | 5 : 0.4407 |
| 6 : 0.4447  | 7 : 0.6029  | 8 : 0.4429  | 9 : 0.4776 |10 : 0.4880 |
|11 : 0.4627  |12 : 0.4593  |13 : 0.2926  |
```

Pour obtenir ces résultats, utiliser :



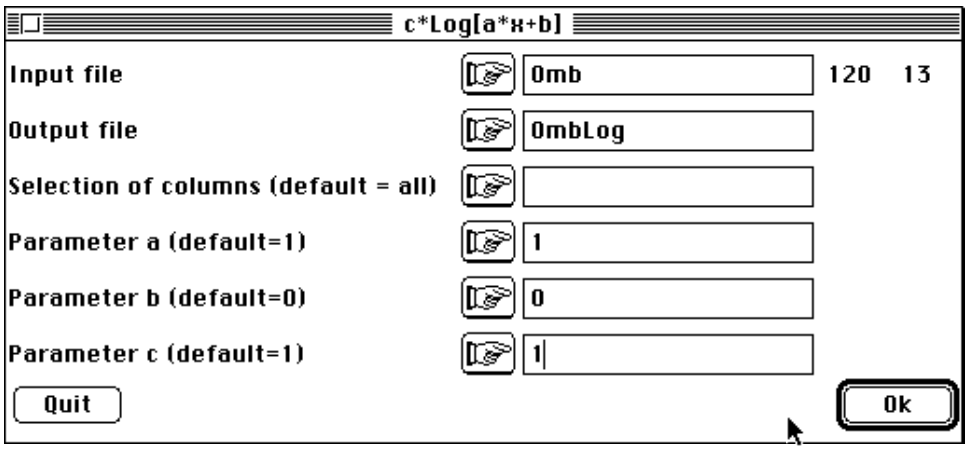
Grossièrement la moitié de la variance est intra-population (53.9%) et l'autre moitié (46.1%) est inter-population. L'effet taille apparaît encore plus nettement sur les plans bivariés. Utilisons par exemple les variables 1 - longueur totale et 7 - diamètre de l'œil (figures 4 et 5).

Il apparaît dans les figures 4 et 5 la redondance propre à l'effet taille. Les deux variables sont corrélées fortement et la séparation de la sous-population 5 y est exprimée deux fois tandis que la différence éventuelle entre les groupes 1-4 (Ain amont/aval indistincts) et 2-3 (Bienne/Loue) y prend une importance limitée.

On peut comparer avec la même représentation obtenue sur les variables normalisées (figure 6). Mesurer les distances entre points ou entre centres des classes sur la figure 6 revient à utiliser la norme de MAHALANOBIS (prise au sens de l'inverse de la matrice de covariance C^{-}, ce qui est un point de vue).

L'effet taille qui va dominer la discussion invite alors à transformer les données en y = Log(x) (Cf. les discussions approfondies dans YOCCOZ 1988, op. cit. , chapitre 3 - L'allométrie : analyses inter-classes et intra-classes et chapitre 4 - Variation géographique : unification des solutions pour la séparation taille-forme).

La discussion univariée et bivariée peut être reprise après cette transformation comme dans la figure 7. Le chapitre 3 porte exclusivement sur le fichier des données transformées par y = Log(x). Dans Bin->Bin :



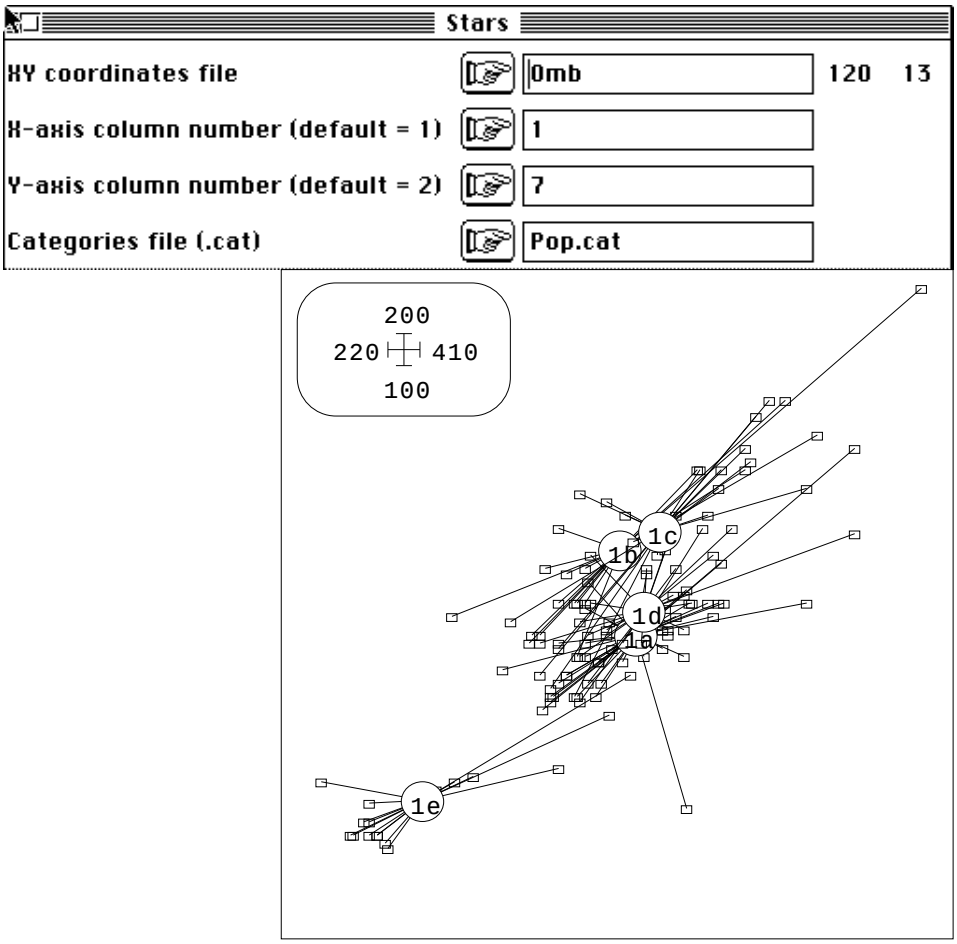


Figure 4 — Nuage bivarié des variables 1 (abscisse) et 7 (ordonnée). Répartition des valeurs entre les 5 sous-populations. Représentation par étoiles dans ScatterClass.

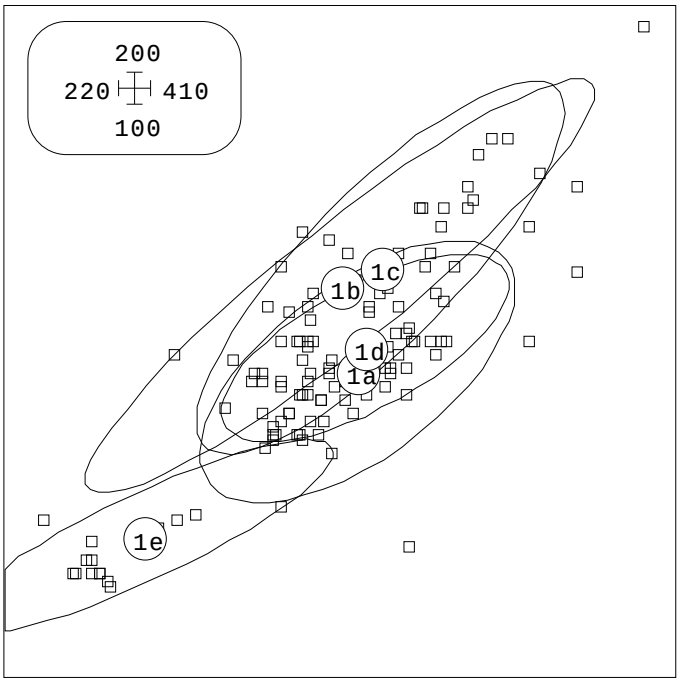


Figure 5 — Nuage bivarié des variables 1 (abscisse) et 7 (ordonnée). Répartition des valeurs entre les 5 sous-populations. Représentation par ellipses de densité (95%).

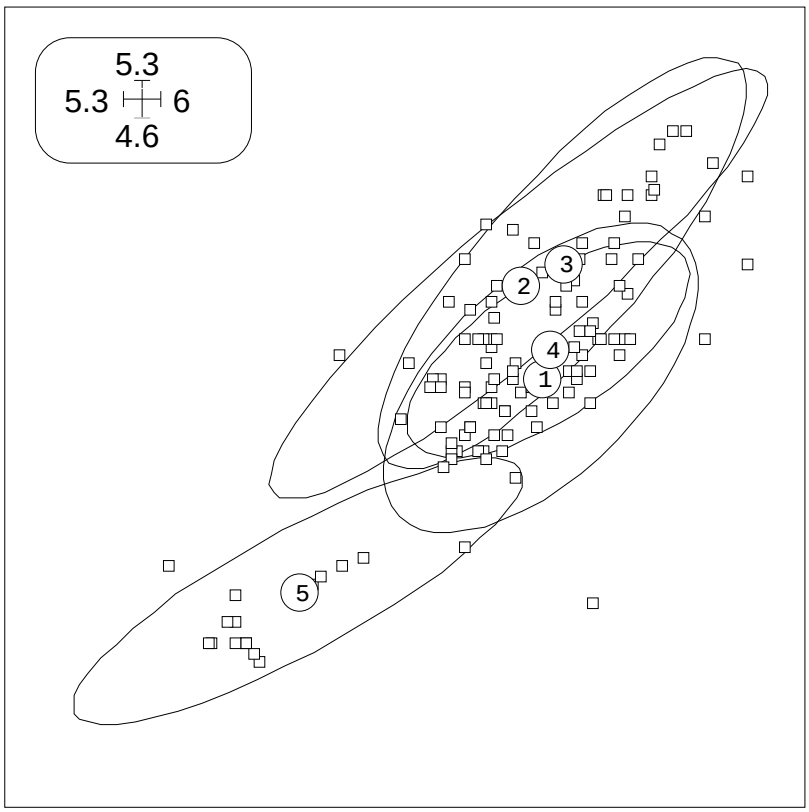


Figure 6 — Nuage bivarié des variables 1 (abscisse) et 7 (ordonnée). Répartition des valeurs entre les 5 sous-populations. Représentation par données transformées en Log.

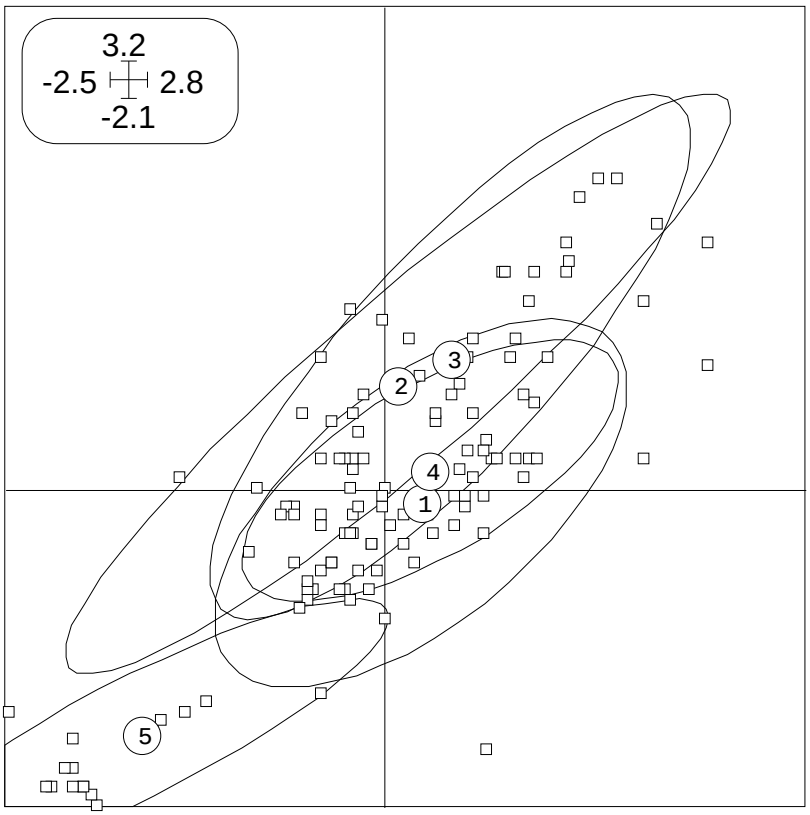


Figure 7 — Nuage bivarié des variables 1 (abscisse) et 7 (ordonnée). Répartition des valeurs entre les 5 sous-populations. Représentation par données en Log normalisées.

3 — Omniprésence de l’effet taille

3.1 — Effet taille et ACP normée

La figure 8 résume l'ACP normée sur les données transformées.

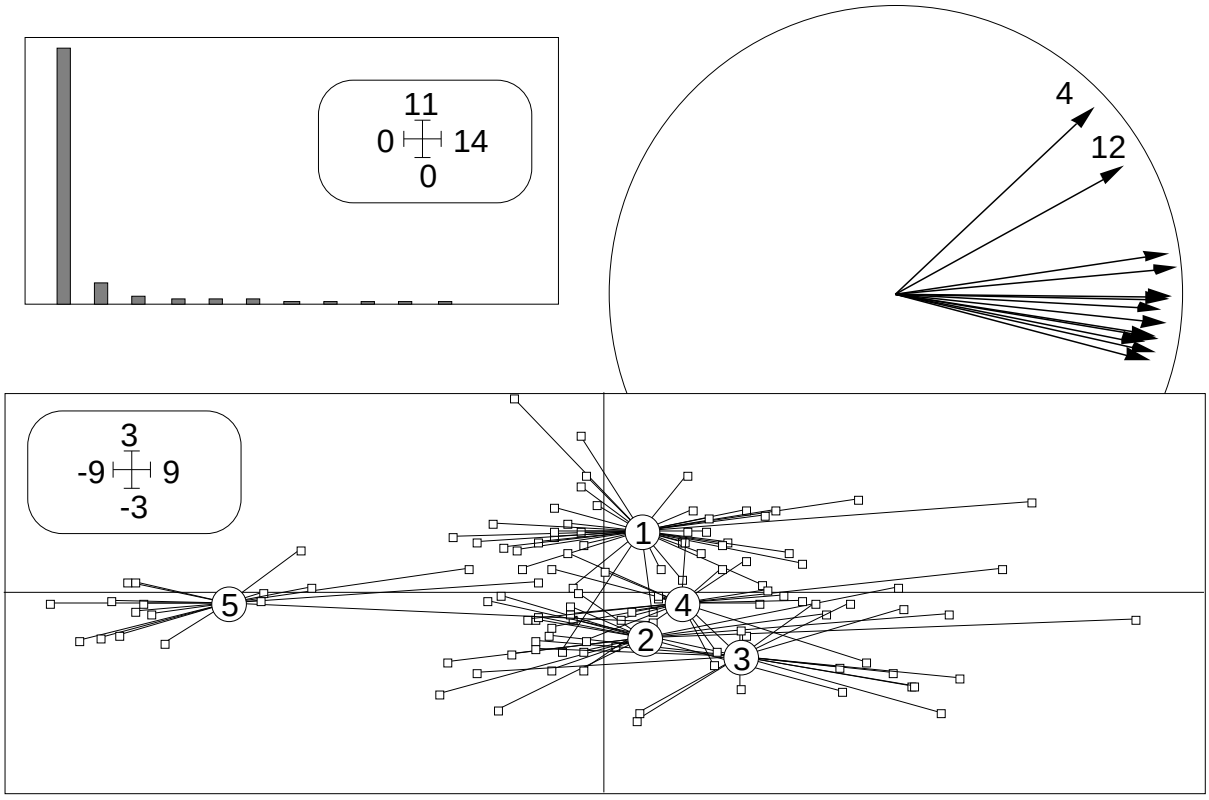


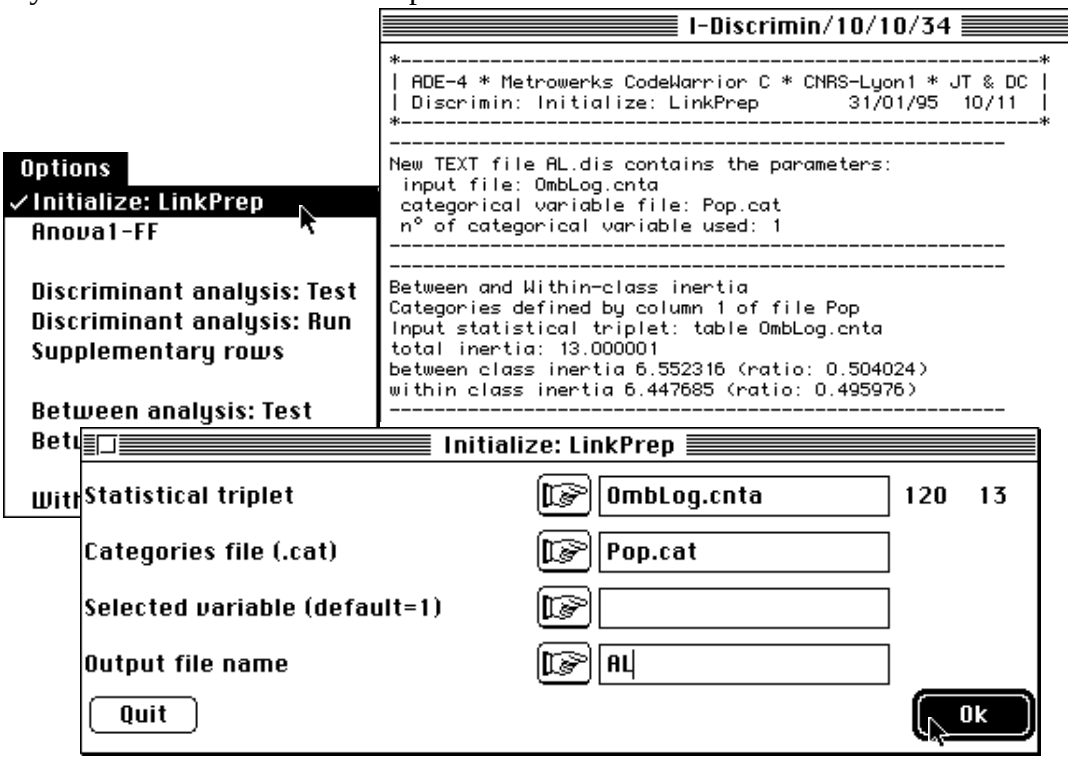
Figure 8 — Résultats graphiques de l'ACP normée des données transformées par y=Log(x).

```
---------------------- Correlation matrix --------------------------
[  1] 1000
[  2]  959 1000
[  3]  954  930 1000
[  4]  739  726  632 1000
[  5]  903  865  892  574 1000
[  6]  876  827  868  499  909 1000
[  7]  837  813  834  493  843  805 1000
[  8]  874  846  882  518  829  813  827 1000
[  9]  903  873  903  606  874  838  838  837 1000
[ 10]  880  845  860  542  856  851  862  831  836 1000
[ 11]  942  910  919  641  876  884  819  845  880  861 1000
[ 12]  800  796  765  716  707  634  596  665  701  659  720 1000
[ 13]  853  749  821  533  812  814  778  787  792  800  822  592 1000
--------------------------------------------------------------------
```

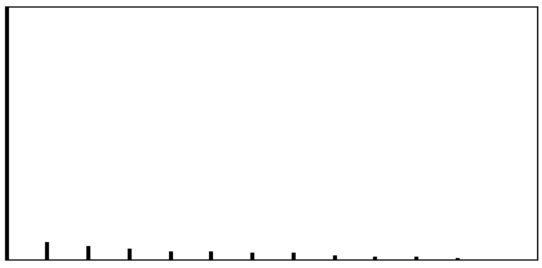
80 % de la variabilité est consacrée à la corrélation entre toutes les variables (effet taille). Les individus de la sous-population 5 sont de taille inférieure aux individus des autres sous-populations pour toutes les variables et la corrélation inter-variables offrent une remarquable confusion entre corrélation intra-groupes (effet taille intra-populations) et inter-groupes (effet de la sous-population 5).

3.2 — Effet taille intra-population

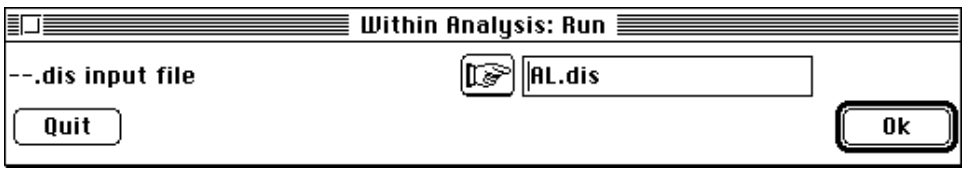
Sur la figure 8, on voit bien que chaque groupe de points propose une ACP dont le premier axe pourrait être l'axe de l'analyse globale. On peut le vérifier par une analyse intra-classes qui recherche l'axe principal du nuage de points formés des sous-nuages par population recentrés. L'analyse intra-classe, comme l'analyse inter-classes ou l'analyse discriminante commence par :



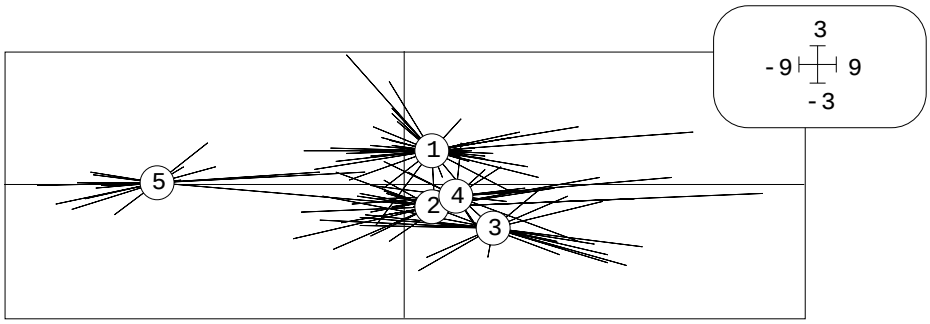
L'inertie totale de 13 (13 variables normalisées) se décompose en deux parties complémentaires. L'inertie intra-classes vaut 6.45 (6.4477 comme inertie totale de l'analyse intra) dont 75% (première valeur propre de l'analyse = 4.766) s'exprime sur un axe :



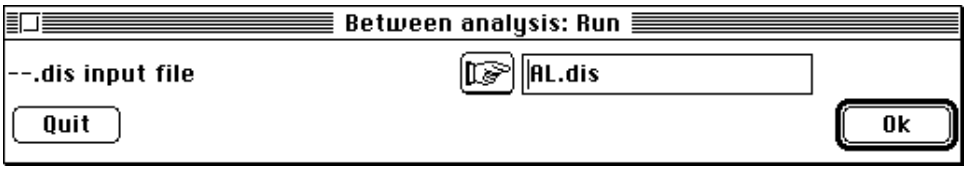
Pour obtenir ce résultat exécuter :



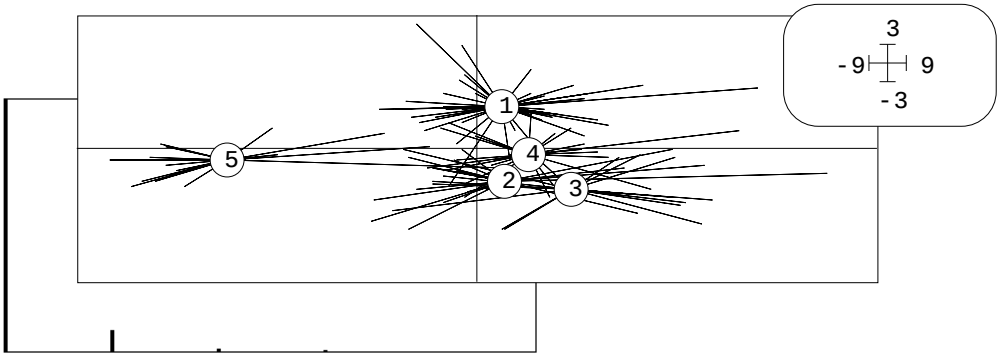
Les inerties projetées sur les axes suivants (5.32%, 4.06%, 3.30%, 2.51%, ...) sont typiques d'une structure aléatoire. La structure intra-classes est un effet taille pur :



Enchaîner avec :



L'inertie inter-classes vaut 6.55 (6.5523 comme inertie totale de l'analyse inter) dont 90% (première valeur propre de l'analyse = 5.918) s'exprime sur un axe. Le second axe est peut-être interprétable.



Ces figures ne sont pas erronées. En haut le nuage est projeté sur le plan optimisant la variance moyenne par groupe des coordonnées et en bas le nuage est projeté sur le plan optimisant la variance des coordonnées des centres de gravité. Le résultat est le même.

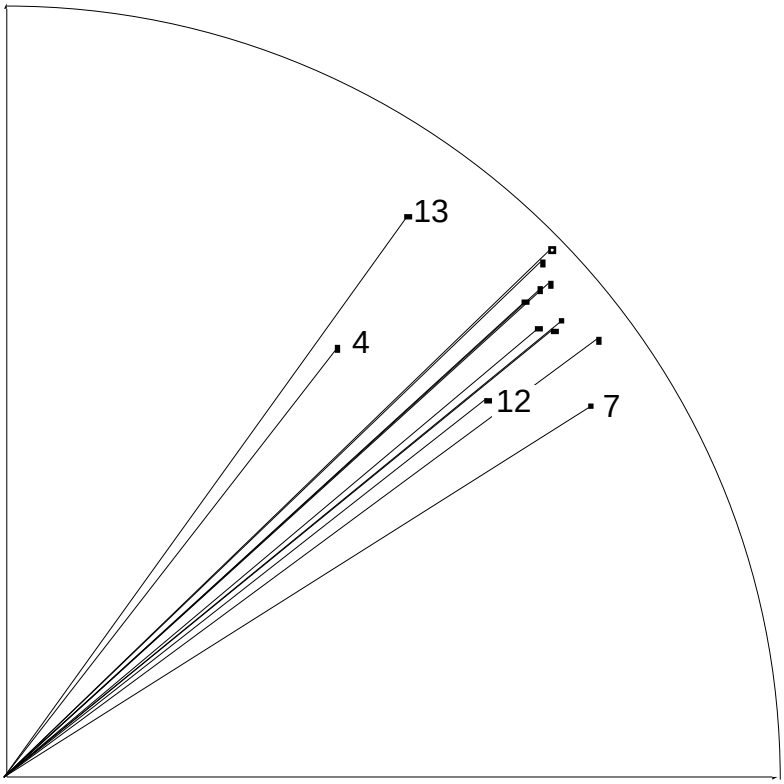


Figure 9 — Projections des variables sur le plan F1 (intra-classes) - F1 (inter-classes). L'effet taille vu comme l'axe F1 de l'ACP normée se partage pour moitié entre structures intra-populations et structures inter-populations.

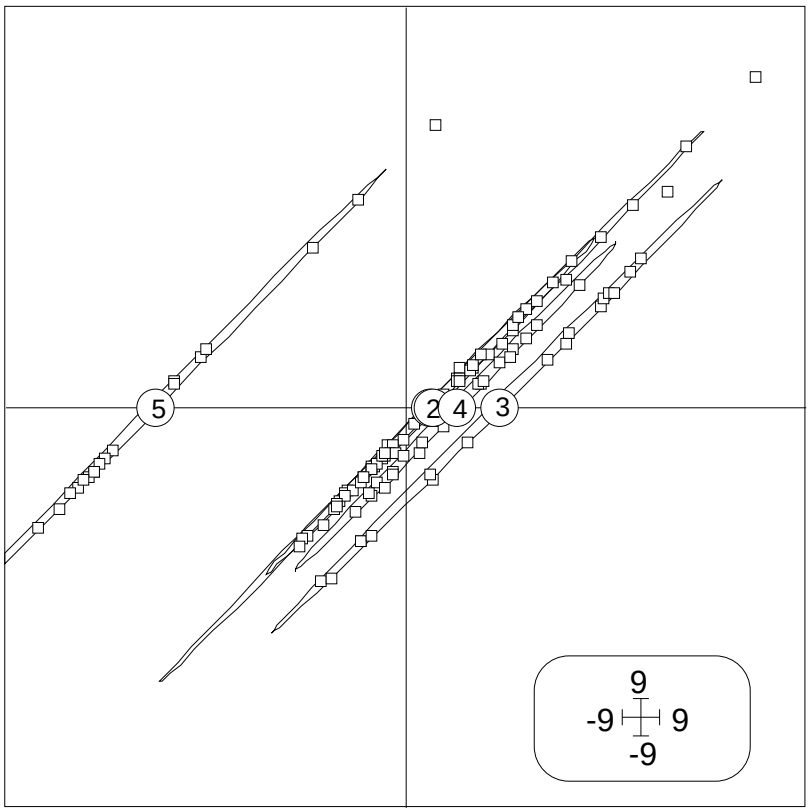


Figure 10 — Positions des individus par leur coordonnées dans l'analyse inter-classes (en abscisse) et leur coordonnée dans l'analyse intra-classes (en ordonnée). L'effet taille intra-populations est parfaitement identique d'une population à l'autre.

L'axe 1 de l'analyse simple, en optimisant la variance, optimise la variance intra et optimise la variance inter. C'est, le moins qu'on puisse dire, un cas rare.

Le plan F1-intra / F1-inter représente une inertie de 10.68 (la première valeur propre de l'ACP normée vaut 10.59) sur une base orthonormée d'un sous-espace de dimension 2 dans $\mathbb{R}^{120}$. Dans cet espace, la première composante principale inter et la première composante principale intra sont deux vecteurs orthogonaux qui définissent un plan sur lequel on peut projeter les variables (figure 9). L'axe 1 de l'ACP simple est sur la bissectrice des composantes inter et intra.

La prépondérance de la corrélation entre toutes les variables induite d'une part par la présence de petits individus dans la population 5 et d'autre part par l'allométrie identique d'une population à l'autre montre combien la décorrélation introduite par l'analyse discriminante est essentielle dans les problèmes de morphométrie.

3.3 — Effet taille et analyse discriminante

La morphométrie a inventé la norme dite de MAHALANOBIS[5]. Le terme a plusieurs sens, qui ici se confondent quasiment. Il s'agit essentiellement de voir que, lorsqu'on mesure la distance entre deux individus avec la norme euclidienne naturelle, on calcule (x_i et y_i sont les mesures sur la variable i) :

$$d^2(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} (x_i - y_i)^2$$

Si 13 variables sont corrélée à la taille une différence de taille est enregistrée 13 fois. Pour se débarrasser de cet artefact (et avoir l'occasion de mesurer autre chose), on peut faire "comme sur la figure 11", c'est-à-dire utiliser la métrique naturelle en remplaçant les variables non par les coordonnées de l'ACP (cela conduit exactement au même résultats) mais par les coordonnées normalisées. Si $\mathbf{U}$ sont les axes principaux et Λ les valeurs propres, on utilise alors la métrique $\mathbf{U}\Lambda^{-1}\mathbf{U}^{\mathbf{t}}$ qui est un inverse ou un inverse généralisé (de rang minimum) de la matrice de corrélation $\mathbf{R}$. On effectue le calcul (L_k et M_k sont les coordonnées sur l'axe k) :

$$d^2(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} \frac{(L_k - M_k)^2}{\lambda_k}$$

Sur la figure 11 le point B est plus proche de C que de A avec la métrique naturelle mais plus proche de A que de C avec la métrique $\mathbf{R^{-1}} = \mathbf{U}\Lambda^{-1}\mathbf{U}^{\mathbf{t}}$. L'effet est encore bien plus sensible en utilisant non pas 2 dimensions, comme sur la figure 11, mais les 13 dimensions de départ. L'analyse discriminante est l'ACP inter-classes utilisant la métrique $\mathbf{R^{-1}}$. On peut espérer que l'effet taille aura donc un effet moins prédominant en utilisant cette métrique:

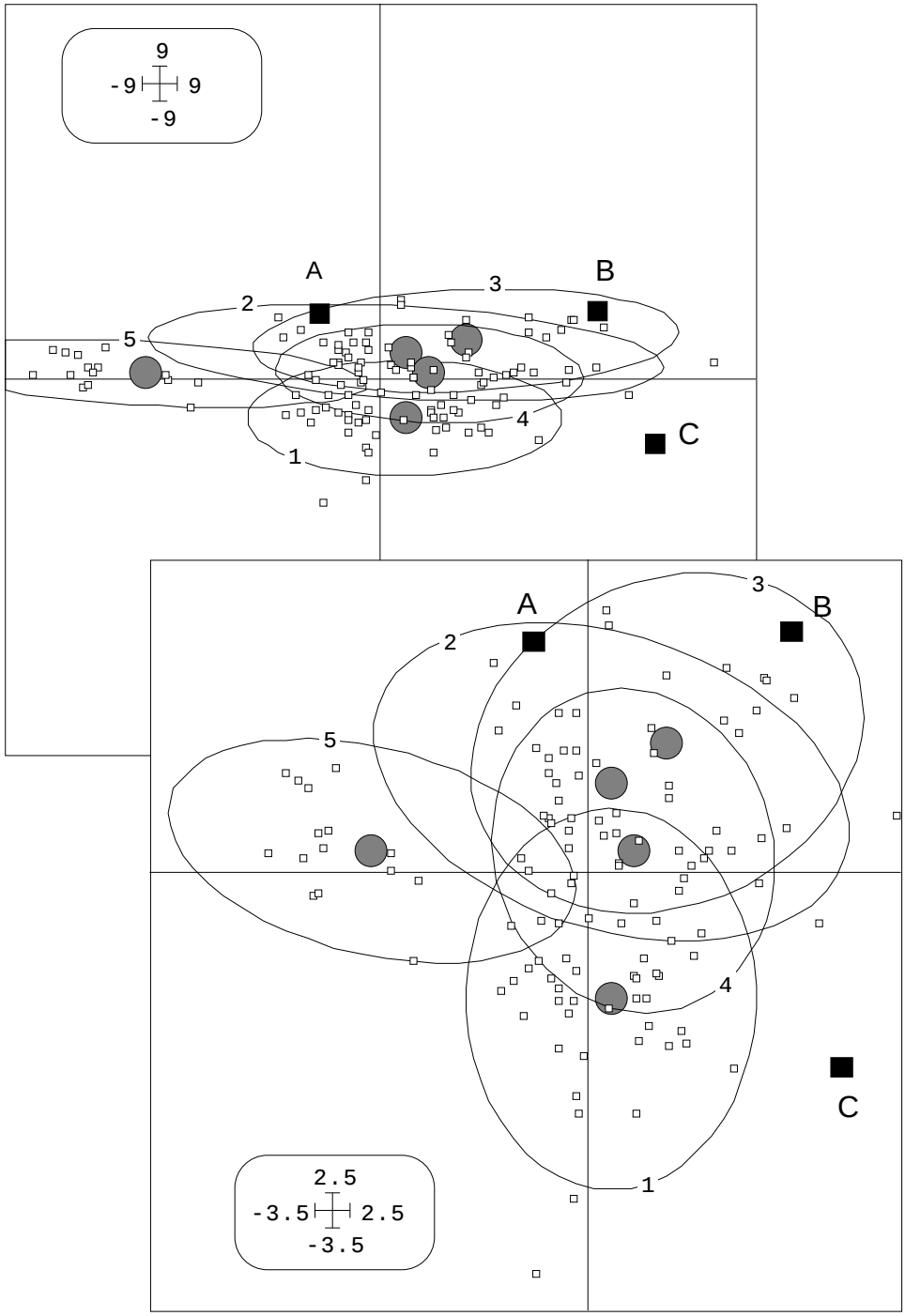
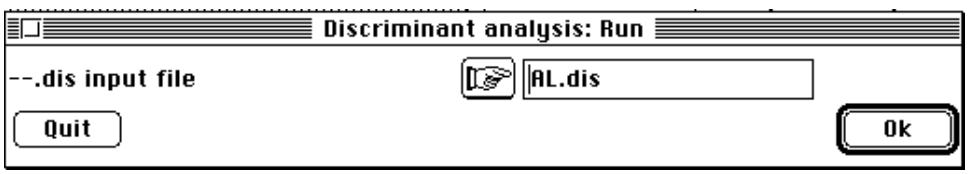


Figure 11 — Plan 1-2 de l'ACP : en haut, coordonnées de variance λ_k (valeur propres) ; en bas, coordonnées de variance 1.

Comme il y a plusieurs variantes dans les programmes d'analyse discriminante on précise les fonctions du module d'ADE.




```
File AL.dima contains the parameters:
   input file: OmbLog.cnta
```

```
   categorical variable file: Pop
   n° of categorical variable used: 1

Discriminant analysis
Categories defined by column 1 of file Pop
Input statistical triplet: table OmbLog.cnta
Number of rows: 120, columns: 13
total inertia (norm C- generalised inverse) = rank of the data matrix: 13.000000

between-class inertia (norm C-): 2.034631 (ratio: 0.156510)
```

Il y a 13 variables et 5 classes. Le programme diagonalise dans R^5 et il ne peut y avoir que 4 valeurs propres non nulles. L'inertie totale vaut 13 car avec la norme de Mahalanobis l'inertie projetée vaut 1 dans chacune des directions. Le taux d'inertie inter-classe a sensiblement diminué.

```
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +7.7336E-01 +0.3801 +0.3801   |02   +6.9364E-01 +0.3409 +0.7210 |
03   +4.2331E-01 +0.2081 +0.9291   |04   +1.4433E-01 +0.0709 +1.0000 |
05   +0.0000E+00 +0.0000 +1.0000

File AL.divp contains the eigenvalues and relative inertia for each axis. It has
5 rows and 2 columns
```

Il y a 13 variables et 5 classes. Le programme diagonalise dans R^{13} et il ne peut y avoir que 4 valeurs propres non nulles. La dernière est nulle. Le schéma utilisé par le programme est du type ACP (**X**, $\mathbf{C}^-$, **D**) ou **X** est le tableau des moyennes par classe et par variable, $\mathbf{C}^-$ une inverse généralisé de la matrice de corrélation (obtenu par diagonalisation : **C** n'est pas nécessairement inversible), **D** la pondération des classes (diagonale des poids des individus). L'inertie totale est la trace de $\mathbf{BC}^-$ dans la notation habituelle (**B** pour between, **W** pour within, **C** pour totale, **C** = **B** + **W**). Les valeurs propres sont des rapports de corrélation (compris entre 0 et 1). Ils s'agit des pourcentages de variance expliquée optimaux qu'on puisse obtenir avec des combinaisons linéaires des variables de départ (sous contrainte de non corrélation progressive). Ici on sait qu'il y a des différences significatives entre populations. Les deux premières valeurs propres voisines (77% et 69%) sont conservées.

```
File AL.dicp contains the correlations between PCA scores
and DA scores. It has 13 rows and 2 columns
File :AL.dicp
----------------------Minimum/Maximum:
Col.:   1  Mini =  -0.75553   Maxi =  0.22479
Col.:   2  Mini =  -0.79753   Maxi =  0.36362
```

Ce fichier contient les coefficients de corrélation entre les variables canoniques (combinaisons linéaires des variables normalisées de départ de variance totale unité et de variance inter-classes maximales) et les coordonnées des individus dans l'ACP normée de départ. On peut ainsi repérer éventuellement l'occurrence de facteurs (ACP) lointains et peu fiables (ci-dessous, à gauche).

```
File AL.diax contains the principal axes
It has 13 rows and 2 columns
File :AL.diax
----------------------Minimum/Maximum:
```

```
Col.:   1  Mini =  -0.89709   Maxi =  -0.22942
Col.:   2  Mini =  0.0090673  Maxi =  0.68237
```

Ce fichier contient les coefficients de corrélation entre les variables canoniques (combinaisons linéaires des variables normalisées de départ de variance totale unité et de variance inter-classes maximale) et les variables de départ (ci-dessous, à droite).

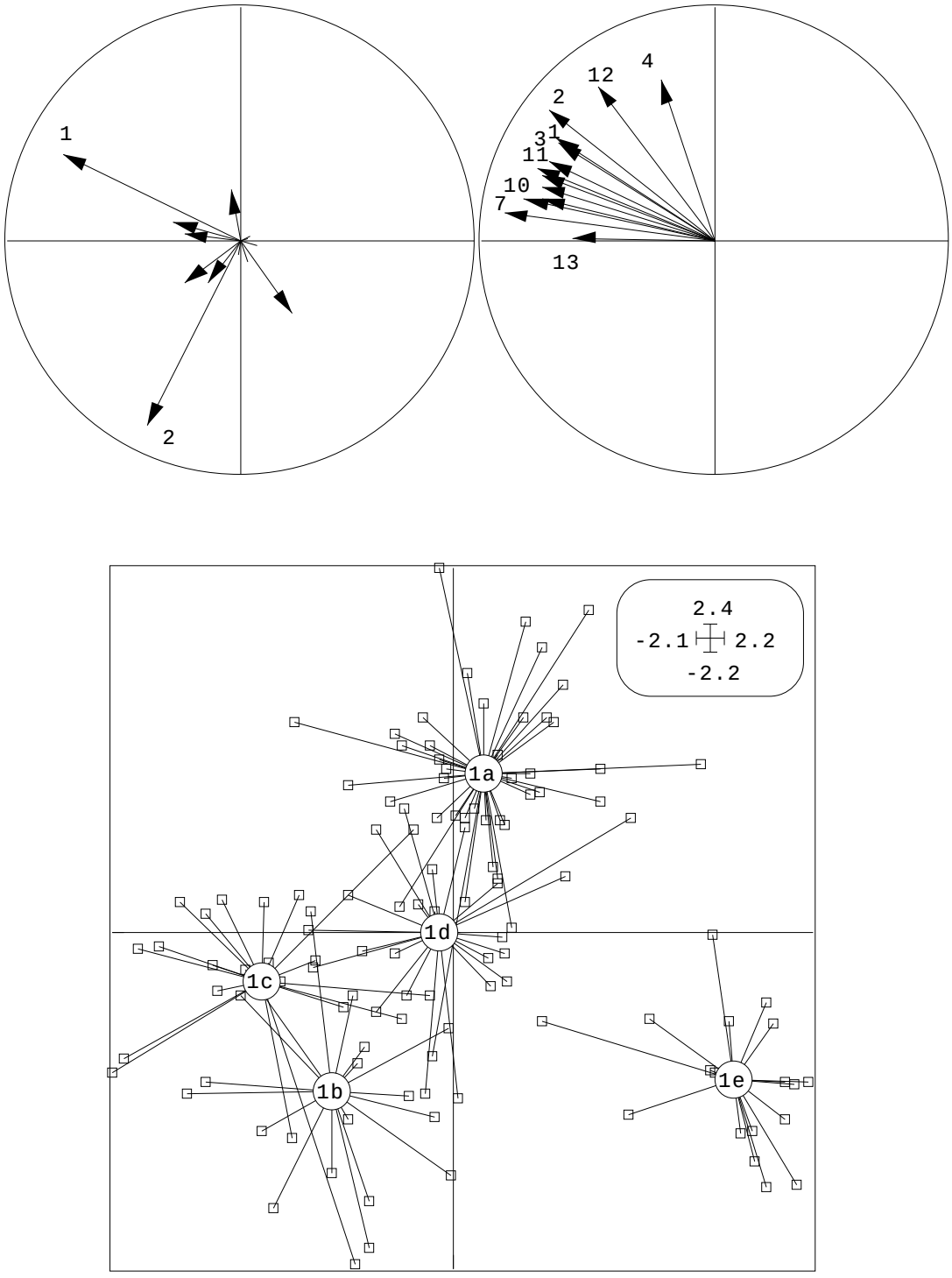


Figure 12 — Plan 1-2 de l'analyse discriminante. Représentation du critère optimisé par des étoiles.

```
File AL.difa contains coefficient of discriminant scores
```

```
It has 13 rows and 2 columns
File :AL.difa
----------------------Minimum/Maximum:
Col.:   1  Mini =  -0.79557   Maxi =  0.83959
Col.:   2  Mini =  -0.67806   Maxi =  0.9208
```

Ce fichier contient les poids canoniques (coefficients des combinaisons linéaires des variables normalisées de départ de variance totale unité et de variance inter-classes maximales). Il permet de projeter éventuellement des individus supplémentaires. On peut utiliser l'un ou l'autre de ces deux fichiers pour repérer le rôle des variables dans la discrimination. Des incohérences entre les deux représentations invitent à la prudence dans l'interprétation.

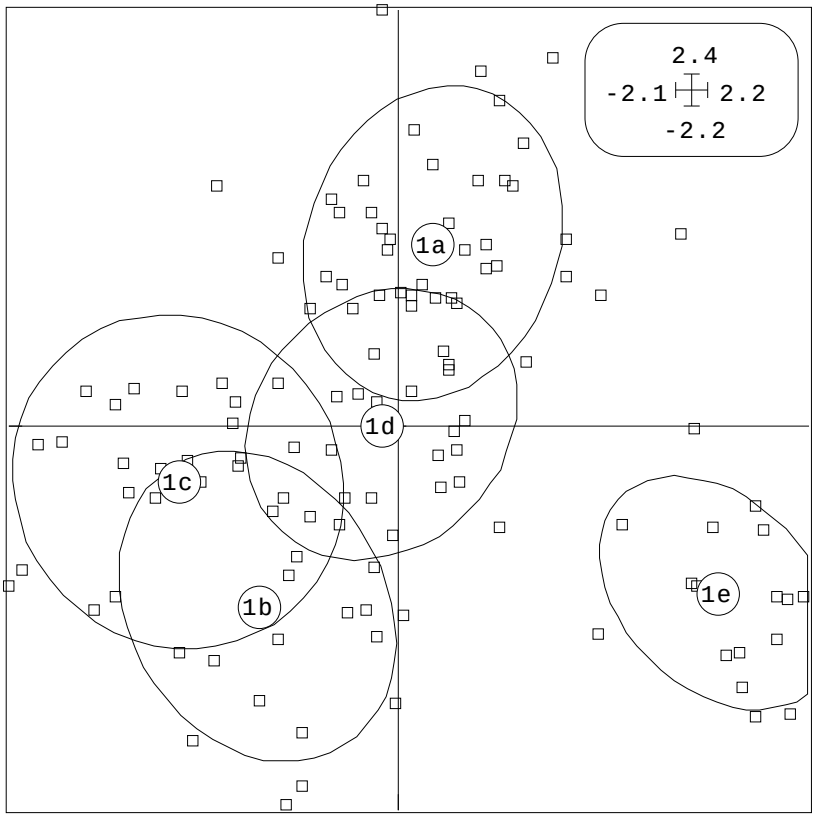


Figure 13 — Plan 1-2 de l'analyse discriminante. Représentation du critère optimisé par des ellipses (70%).

```
File AL.dili contains canonical row scores with unit norm
It has 120 rows and 2 columns

File :AL.dili
----------------------Minimum/Maximum:
Col.:   1  Mini =  -2.0887  Maxi =  2.1768
Col.:   2  Mini =  -2.1717  Maxi =  2.3912
```

Ce fichier contient les valeurs des variables canoniques (combinaisons linéaires des variables normalisées de départ de variance totale unité et de variance inter-classes maximales).

Les cartes factorielles des individus peuvent utiliser un des procédés graphiques des figures 12 à 15. Suivant les cas on sera amené à préférer l'un ou l'autre de ces systèmes de représentation.

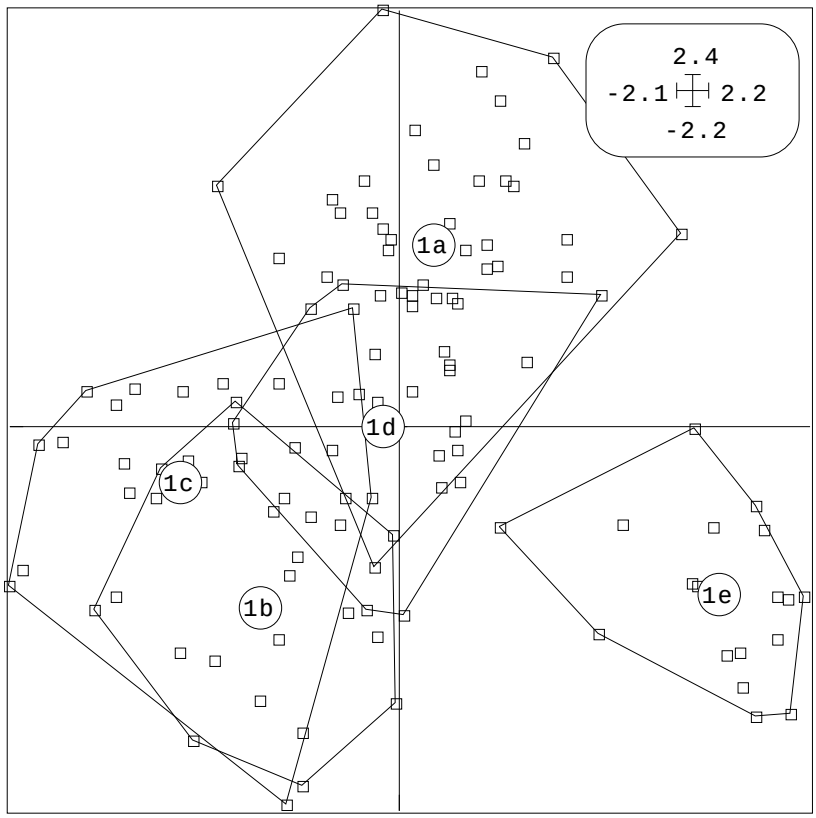
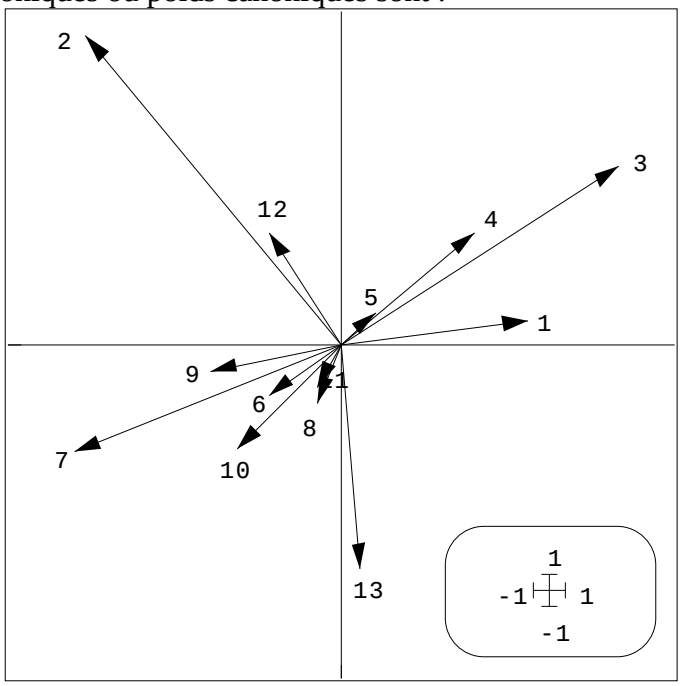


Figure 14 — Plan 1-2 de l'analyse discriminante. Représentation du critère optimisé par des polygones de contour apparent par groupes.

On remarque que le plan 1-2 de l'ACP joue le rôle principal ce qui confirme l'interprétabilité du plan 1-2 discriminant. On retrouve la liaison F1-ACP et l'ensemble des variables dans la bissectrice du plan 1-2 de l'analyse discriminante. Mais l'effet taille qui isole le groupe 5 ne joue plus un rôle écrasant. Les coefficients des variables dans les codes canoniques ou poids canoniques sont :



On trouvera une interprétation détaillée dans YOCCOZ (1988 op. cit. p. 127 et suivantes) et dans PERSAT (1988*, p.53 et suivantes). Dans la figure qui précède réside la principale difficulté d'interprétation. Listons les valeurs utilisées :

1	0.5668	0.0704
2	-0.7615	0.9208
3	0.8396	0.5288
4	0.4051	0.3320
5	0.1084	0.0928
6	-0.2159	-0.1555
7	-0.7956	-0.3258
8	-0.0673	-0.1793
9	-0.3868	-0.0815
10	-0.3103	-0.3191
11	-0.0644	-0.1298
12	-0.2090	0.3357
13	0.0638	-0.6781

La variable canonique 1 s'écrit :

$f(M) = 0.567\ x_1 - 0.761\ x_2 + 0.840\ x_3 + \ldots - 0.209\ x_{12} + 0.064\ x_{13}$

où x_j est la valeur normalisée de la variable j pour un individu. La variable 2 a dans cette combinaison un coefficient négatif et la variable 3 un coefficient positif. Or le résultat est corrélé positivement avec les 2 variables. Toute interprétation rapide des coefficients est périlleuse (TOMASSONE & Coll 1988[6] p. 64). Dans les deux représentations les variables 2 et 7 ont une position cohérente sur les 2 axes. C'est pourquoi on retiendra pour retourner aux données la figure 15. Pour en savoir beaucoup plus sur l'analyse discriminante classique, utiliser la référence qui précède.

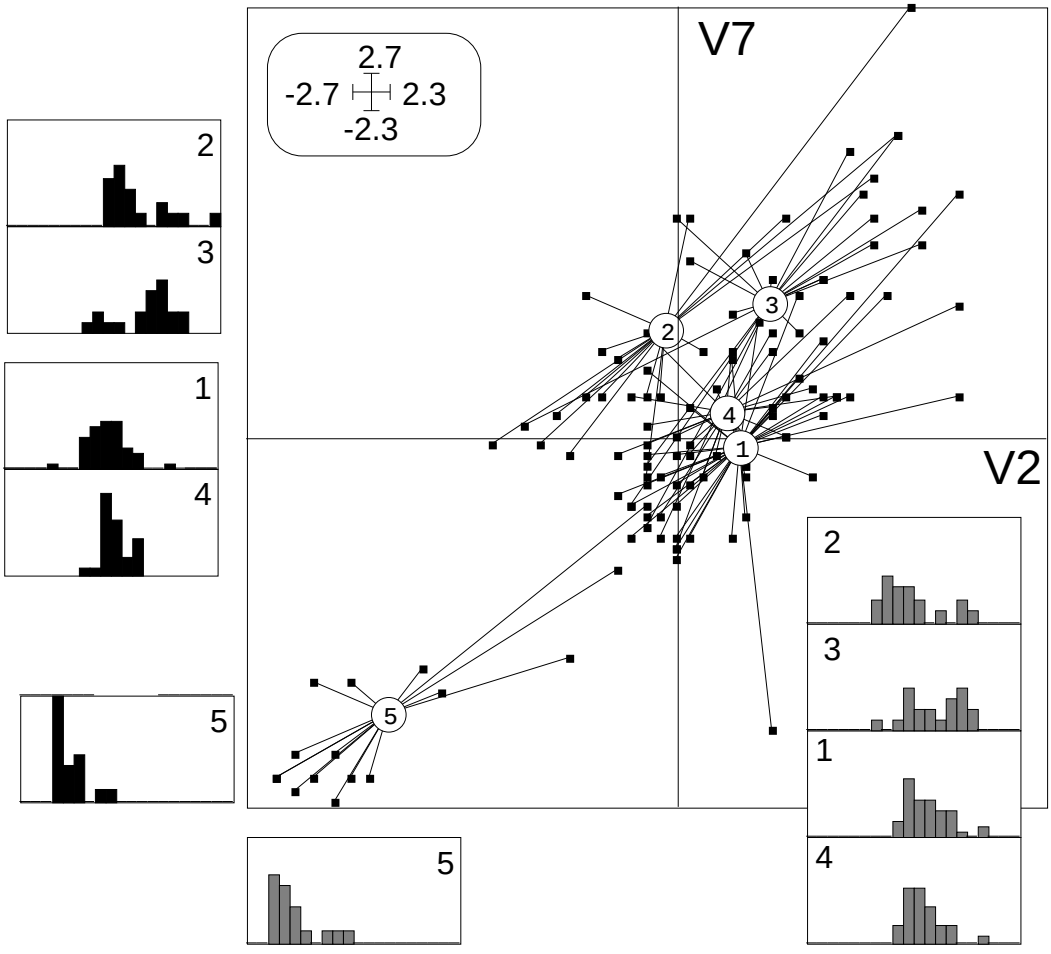


Figure 15 — Illustration de la capacité discriminante du couple des variables 2 (V2, distance prédorsale) et 7 (V7, diamètre de l'œil). Données en Log normalisées. Nuage bivarié et représentation des 5 populations. En gris : histogrammes par population de la variable 2 : séparation du groupe 5. En noir : histogrammes par population de la variable 7 : séparation du groupe 5 et distinction entre les groupes 2-3 et 1-4. La majeur partie de l'information contenue dans le plan discriminant (figures 10 à 13) est exprimée sur cette relecture élémentaire des données.

Le caractère globalement plus petit des individus de la population 5 est resté l'élément principal des analyses en composantes principales normée, inter et intra classes et discriminante classique. L'élimination de l'effet taille ne peut être, ici, qu'intentionelle.

4 — Analyses de la forme

Nous pouvons faire deux essais pour éliminer délibérément l'effet taille. On se reportera à l'article de YOCCOZ (1993, op. cit.) pour les explications et références bibliographiques nécessaires.

4.1 — Double centrage additif sur les logarithmes

Le premier est celui de l'ACP doublement centrée (en ligne et en colonne) sur le fichier transformé par $y = \mathrm{Log}(x)^7$. La figure 15 résume cette analyse qui n'autorise plus la prise en compte d'un effet taille mais montre encore une bonne part de variance inter-classe.

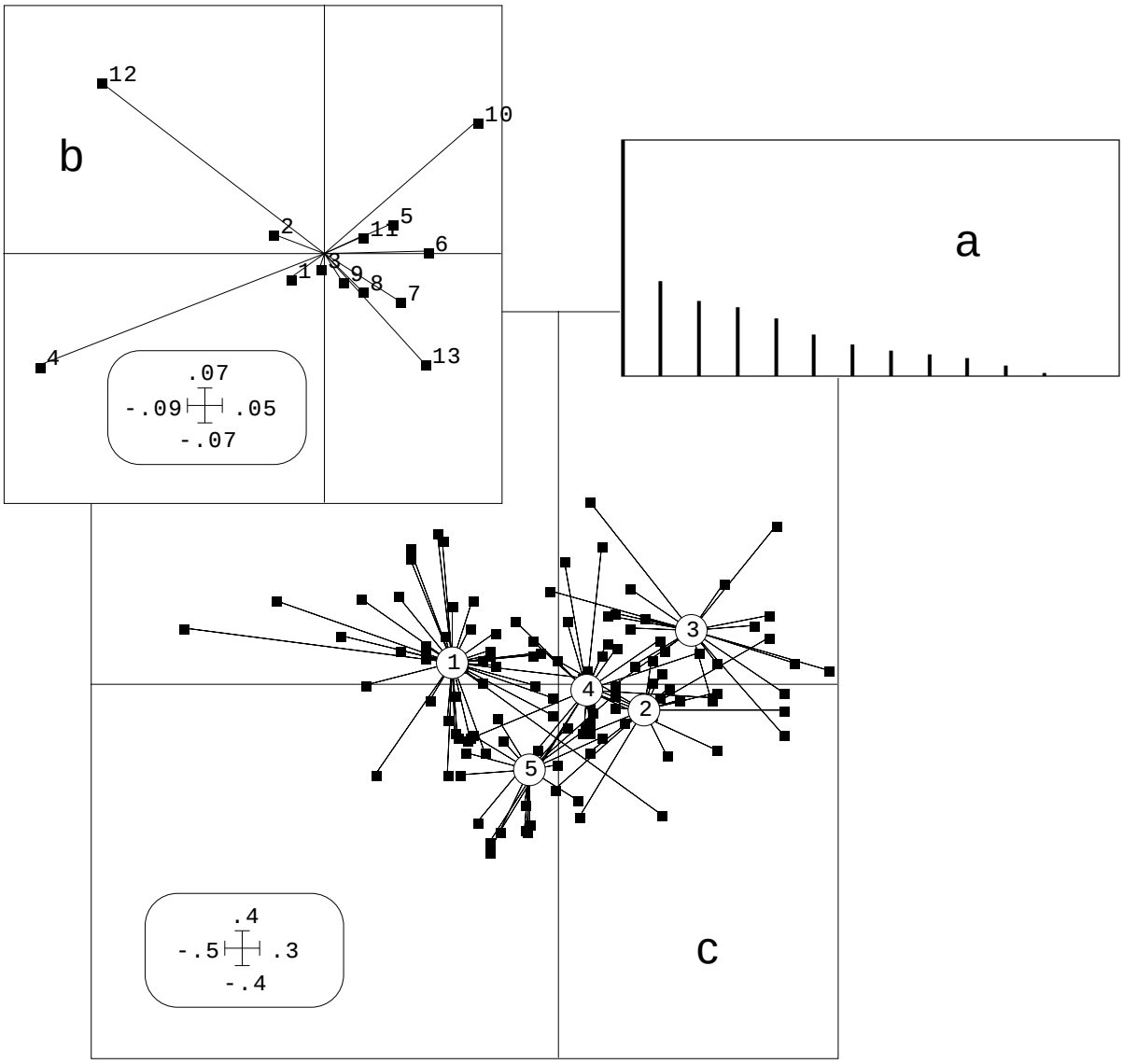


Figure 16 — ACP du tableau des logarithmes doublement centrés. a - graphe des valeurs propres. b - carte des variables. c- cartes des individus. Les deux cartes factorielles sont centrées (in Yoccoz 1988 op. cit. p. 126).

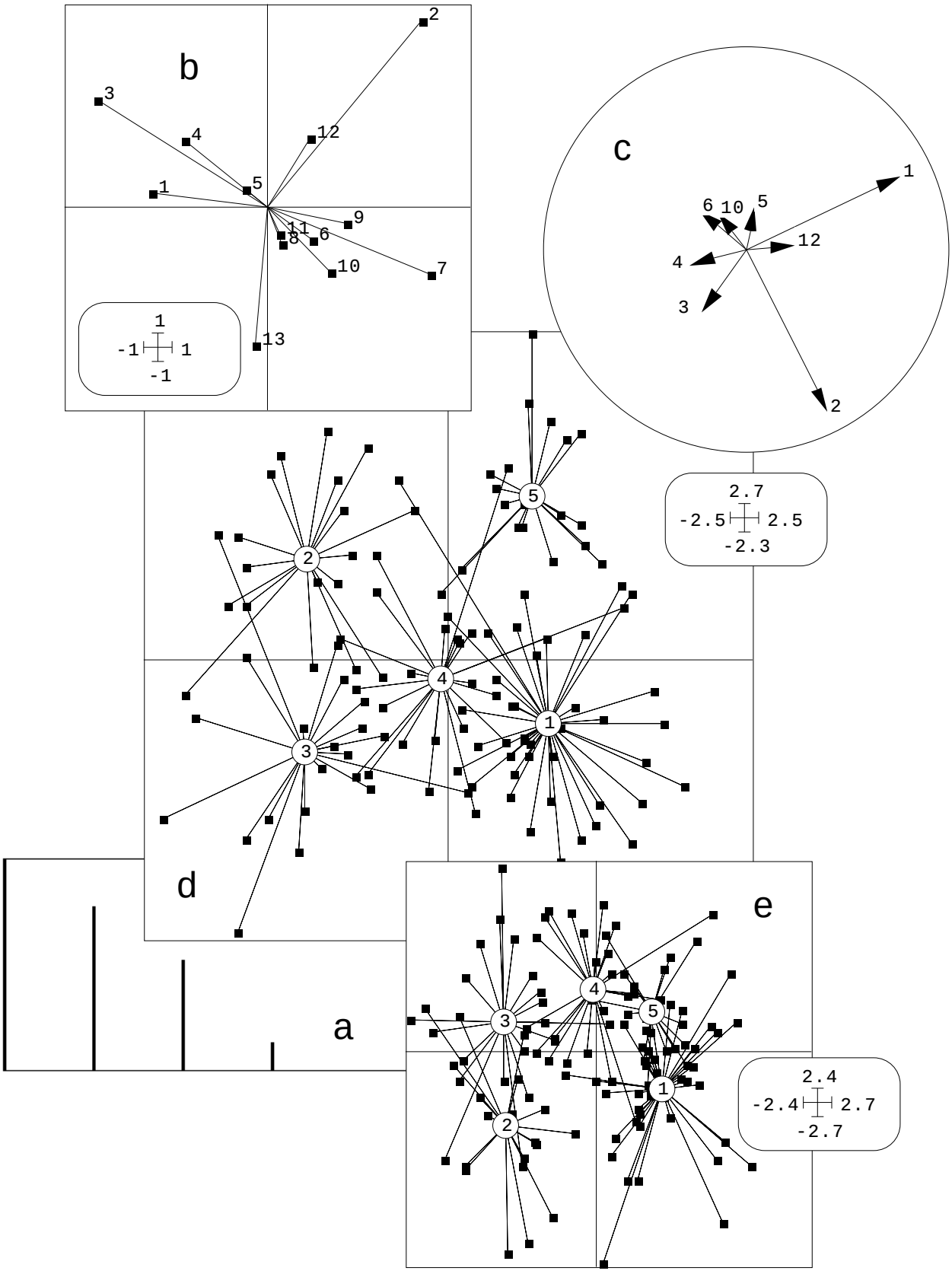
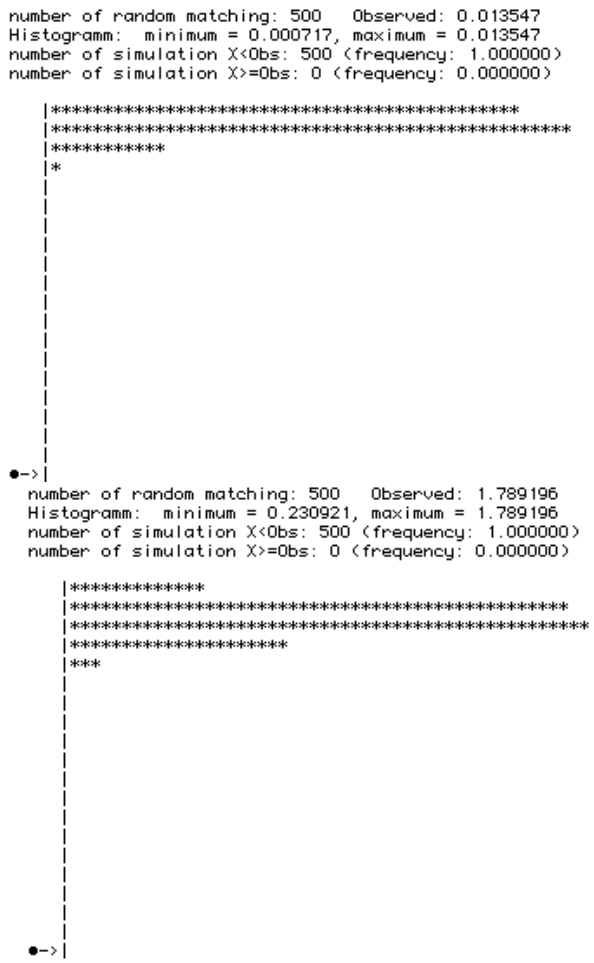


Figure 17 — Analyse discriminante sur tableau en Log doublement centré. a - valeurs propres. b - coefficients des combinaisons linéaires optimales (poids canoniques). c - corrélations facteurs ACP et facteurs AD. d - carte factorielle des individus (plan 1-2). e - carte factorielle des individus (plan 1-3).

On peut s'interroger sur la signification statistique de la variabilité inter-classes. On utilisera ici un test de permutations sur la trace de l'analyse inter-classes (à gauche) ou de l'analyse discriminante (à droite) :



La variance inter-classe observée est incompatible avec l'hypothèse d'identité des populations, tant avec la norme naturelle qu'avec la métrique inverse associée à l'ACP doublement centrée. La figure 18 résume l'analyse discriminante.

On notera la grande stabilité apparente de la variabilité intra-groupe et la capacité uniforme de distinguer partiellement chaque groupe de chacun des autres. La structure des données, vues de cette manière, est essentiellement une structure inter-groupe, puisque le plan 1-2 de l'analyse discriminante est voisin du plan 1-2 de l'analyse de base.

4.2 — Double centrage multiplicatif sur les données

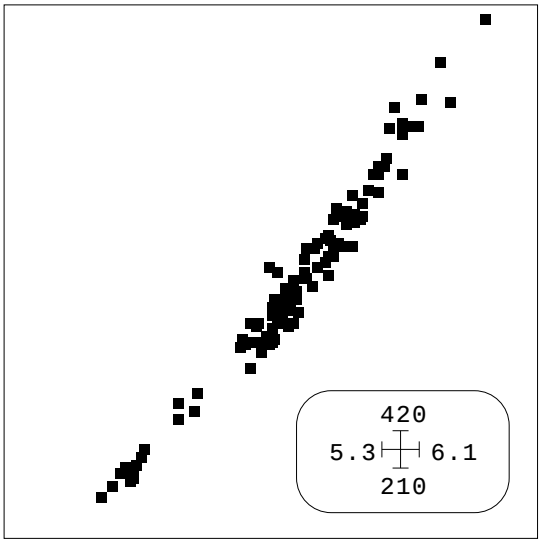
Sans passer par la transformation y = Log(x), on peut directement estimer un modèle de taille en supposant qu'il existe un poisson standard fictif **a** = (a_1, a_2, ...a_{13}) et qu'une observation s'écrit :

$$\mathbf{x} = (x_1, x_2, ...x_{13}) = k\,(a_1, a_2, ...a_{13}) + (y_1, y_2, ...y_{13})$$

Le modèle de taille dérive d'une ACP non centrée dont le premier facteur donne une estimation aux moindres carrés[8]. La valeur relative des six ACP de base d'un tableau homogène justifie pleinement ce point de vue :

```
Option  1 = No action (non centred PCA)            Inertia =   899833
Option  2 = Centred table (overall centred PCA)    Inertia =   316313
Option  3 = Centred (zero mean) columns            Inertia =    14712
Option  4 = Centred (zero mean) rows               Inertia =   308538
Option  5 = Additive model                         Inertia =     6937
Option  6 = Multiplicative model                   Inertia =     1928
```

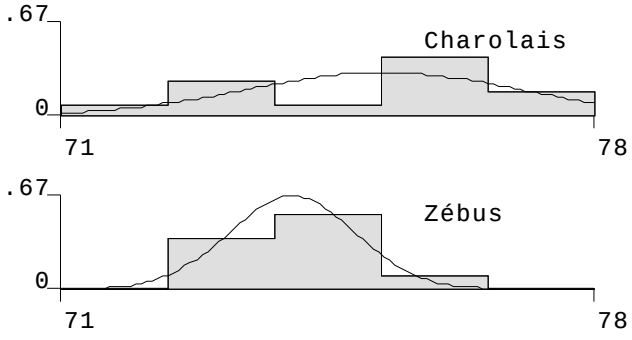
On note aussi la communauté des estimations de la taille faite par la moyenne des valeurs en Log (en abscisse, ci-dessous) et la première composante principale de l'ACP non centrée (en ordonnée) :



Après élimination de l'effet taille de cette manière, l'ACP des résidus conduit à une analyse inter-classes, une analyse intra-classe et une analyse discriminante dont les résultats sont très voisins des précédentes. On veut simplement indiquer par ces remarques que l'analyse discriminante, surtout connue dans le cadre de l'ACP normée, s'étend à tous les types de variables et tous les modes de préparation des données.

On illustre ce dernier point avec l'exemple "Charolais-Zébus" de TOMASSONNE & coll. (op. cit. p. 43). Le tableau 2 reproduit les données et l'estimation de la taille par un modèle multiplicatif.

La taille est-elle un facteur discriminant entre les deux groupes ? La représentation du poids de viande estimée par l'effet taille demande un test de signification :



	A		B		C		D		E		F	
1	395	395.41	224	225.61	35.1	29.49	79.1	73.58	6	8.84	14.9	16.18
2	410	408.87	232	233.29	31.9	30.49	73.4	76.08	9.7	9.14	16.4	16.73
3	405	405.96	233	231.62	30.7	30.27	76.5	75.54	7.5	9.07	16.5	16.61
4	405	408.7	240	233.19	30.4	30.48	75.3	76.05	8.7	9.14	16	16.72
5	390	388.36	217	221.59	31.9	28.96	76.5	72.27	7.8	8.68	15.7	15.89
6	405	410.29	243	234.1	32.1	30.6	77.4	76.35	7.1	9.17	15.5	16.79
7	390	393.62	229	224.59	32.1	29.35	78.4	73.25	4.6	8.8	17	16.11
8	405	408.87	240	233.29	31.1	30.49	76.5	76.08	8.2	9.14	15.3	16.73
9	420	417.37	234	238.14	32.4	31.12	76	77.66	7.2	9.33	16.8	17.08
10	390	391.04	223	223.12	33.8	29.16	77	72.76	6.2	8.74	16.8	16
11	415	418.97	247	239.05	30.7	31.24	75.5	77.96	8.4	9.36	16.1	17.14
12	400	402.96	234	229.91	31.7	30.05	77.6	74.98	5.7	9.01	18.7	16.49
13	400	398.03	224	227.1	28.2	29.68	73.5	74.06	11	8.9	15.5	16.29
14	395	396.65	229	226.32	29.4	29.58	74.5	73.81	9.3	8.87	16.1	16.23
15	395	392.33	219	223.85	29.7	29.26	72.8	73	8.7	8.77	18.5	16.05
16	395	394.44	224	225.05	28.6	29.42	73.7	73.4	8.7	8.82	17.3	16.14
17	400	397.61	223	226.86	28.5	29.65	73.1	73.99	9.1	8.89	17.7	16.27
18	400	397.96	224	227.06	27.8	29.68	73.2	74.05	12.2	8.89	14.6	16.28
19	400	396.53	221	226.25	26.5	29.57	72.3	73.79	13.2	8.86	14.5	16.22
20	410	408.84	233	233.27	25.9	30.49	72.3	76.08	11.1	9.14	16.6	16.73
21	402	403.46	234	230.2	27.1	30.09	72.1	75.08	10.4	9.02	17.5	16.51
22	400	397.16	223	226.61	26.8	29.62	70.3	73.9	13.5	8.88	16.2	16.25
23	400	392.97	213	224.21	25.8	29.31	70.4	73.12	12.1	8.78	17.5	16.08

Tableau 2 — Mesures de 6 variables (poids en kg) sur 12 Charolais et 11 Zébus. Estimation des données par un modèle multiplicatif (une valeur observée est approximée par le produit d'un effet individu et d'un coefficient d'échelle associé à la variable). A - poids vif, B - poids de carcasse, C - poids de viande première qualité, D - poids de viande totale, E - poids de gras, F - poids d'os.

L'analyse de variance répond :

variable 1 from taille versus variable 1 from zcla

Source *	SS*	d.f.*	MS*	F*	Proba*
Between *	.2341E+03*	1*	.2341E+03*	.3690E+01*	0.06842*
Within *	.1332E+04*	21*	.6344E+02*		
Total *	.1566E+04*	22*			

Un test de permutation donne un avis remarquablement identique :

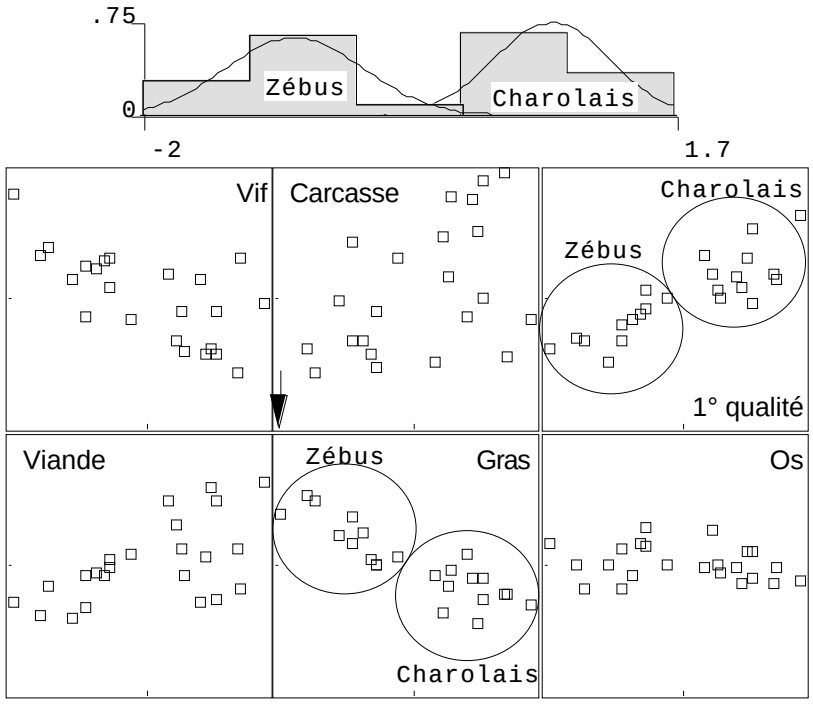
```
number of random matching: 10000   Observed: 0.149455
Histogramm:  minimum = 0.000000, maximum = 0.541711
number of simulation X<Obs: 9341 (frequency: 0.934100)
number of simulation X>=Obs: 659 (frequency: 0.065900)
```

Quant à la forme, elle ne laisse pas planer d'ambiguïté :

```
number of random matching: 1000   Observed: 0.811388
Histogramm:  minimum = 0.008441, maximum = 0.811388
number of simulation X<Obs: 1000 (frequency: 1.000000)
number of simulation X>=Obs: 0 (frequency: 0.000000)

    |*********
    |**************************
    |****************************************
    |**************************************************
    |*********************************************
    |***********************************
    |*************************************
    |*********************************
    |************************
    |************
    |***********
    |******
    |***
    |*
    |*
    |
    |
    |
    |
●->|
```

De l'analyse discriminante sur le tableau des résidus, on retiendra le graphe qui relie le code canonique et chacune des variables et tend à identifier comme principales variables actives le couple 3 et 5 :



On notera donc la possibilité laissée à l'utilisateur de faire de l'analyse discriminante après une quelconque pratique de centrage, en particulier après une AFC. On peut refaire les calculs avec les cartes Ombres et Zébus de la pile ADE•Data.

Références

[1] Persat, H. (1988) De la biologie des populations de l'Ombre commun (Thymallus thymallus (L. 1758)) à la dynamique des communautés dans un hydrosystème fluvial aménagé, le Haut-Rhône français. Eléments pour un changement d'échelles. Thèse d'état. Université Lyon 1. 1-223.

[2] Surre, C., Persat, H. & Gaillard, J.M. (1986) A biometric study of three populations of the European grayling, Thymallus thymallus (L.), from the french Jura mountains. *Canadian Journal of Zoology* : 64, 2430-2438.

[3] Yoccoz, N. (1988) Le rôle du modèle euclidien d'analyse des données en biologie évolutive. Thèse de doctorat, Université Lyon 1. 1-254.

[4] Yoccoz, N. G. (1993) Morphométrie et analyses multidimensionelles. Une revue des méthodes séparant taille et forme. In : *Biométrie et Environment*. Lebreton, J.D. & Asselain, B. (Eds.) Masson, Paris. 73-99.

[5] Mahalanobis, P.C. (1936) On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India* : 12, 49-55.

[6] Tomassone (R.), Danzard (M.), Daudin (J.-J.) & Masson (J.P.) (1988). *Discrimination et classement*. Masson, Paris. 1-173.

[7] Darroch , J.N. & Mosimann, J.E. (1985) Canonical and principal components of shape. *Biometrika* : 72, 241-252.

[8] Whittle, P. (1952) On principal components and least square methods of factor analysis. *Skandinavisk aktuarietidskrift* : 35, 223-239.