Introduction à l'analyse en composantes principales

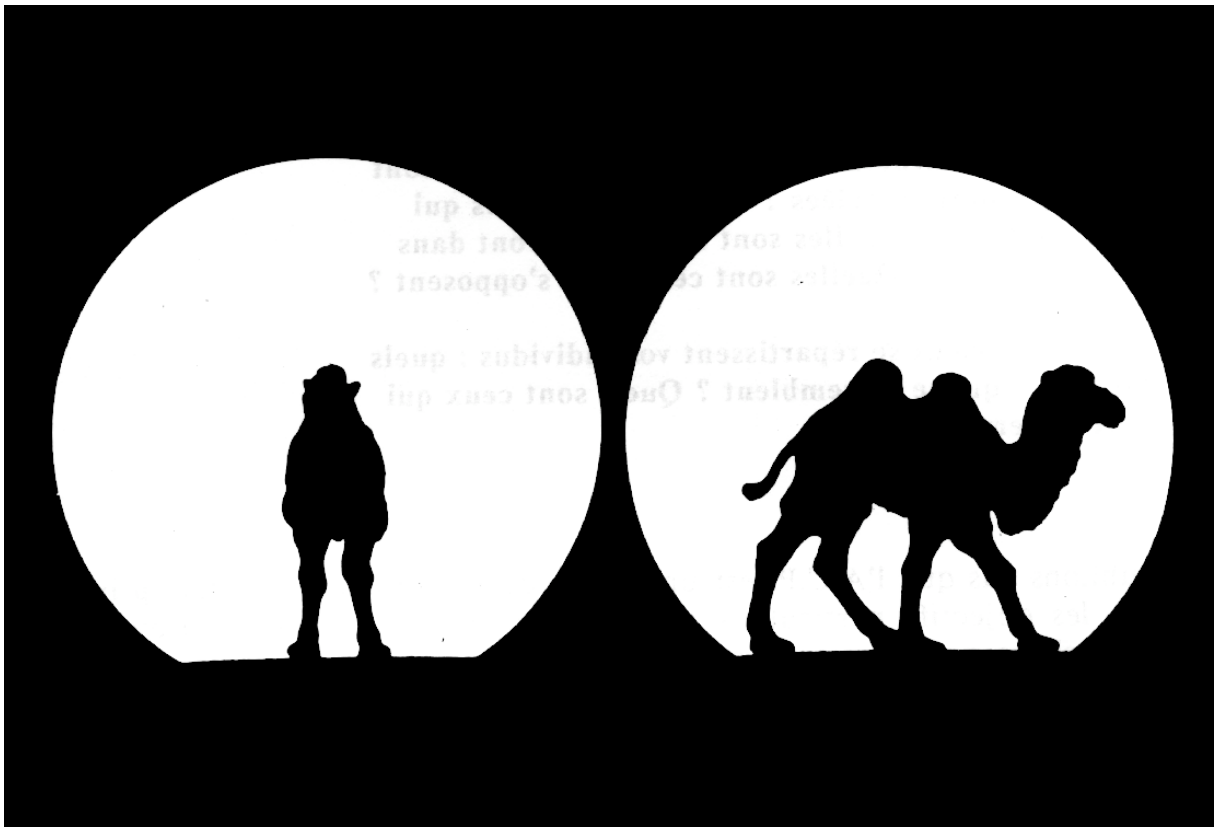
Tables des matières :

Objectifs

L'analyse en composantes principales est la plus ancienne, la moins sophistiquée et la plus centrale des analyses factorielles. L'analyse en facteurs communs et spécifiques , l'analyse des correspondances et canonique des correspondance, l'analyse discriminante en découlent plus ou moins directement. Comprendre les fondements de l'ACP est une porte d'entrée de prédilection dans cette famille d'analyse.

On peut considérer l'ACP comme une analyse descriptive visant à représenter au mieux sur un plan des objets situés dans un hyperespace dont les dimensions échappent à notre contrôle mental et nos possibilités de représentation graphique.

Voici une image très parlante :



L'ombre plane de ce chameau tri-dimensionnel est projetée sur deux plans. Le plan de droite est beaucoup plus informatif que le plan de gauche ! Il permet pratiquement de se passer de la troisième dimension…

Si l'on reprend l'équation : variabilité = information + bruit , trouver le plan le plus informatif revient à trouver les directions dans lesquelles se trouvent la plus grande variance. En effet, si le bruit est plus grand que l'information… les données ne sont guère pertinentes.

L'Analyse en composante principale (ACP)

Cas de deux variables

Imaginons 2 variables standardisées Z1 et Z2 formant la matrice **Z**, et la matrice orthogonale **C** correspondant à une rotation de 45° (voir module 220, rotation orthogonale). Le produit s'écrit

$$Y = ZC$$

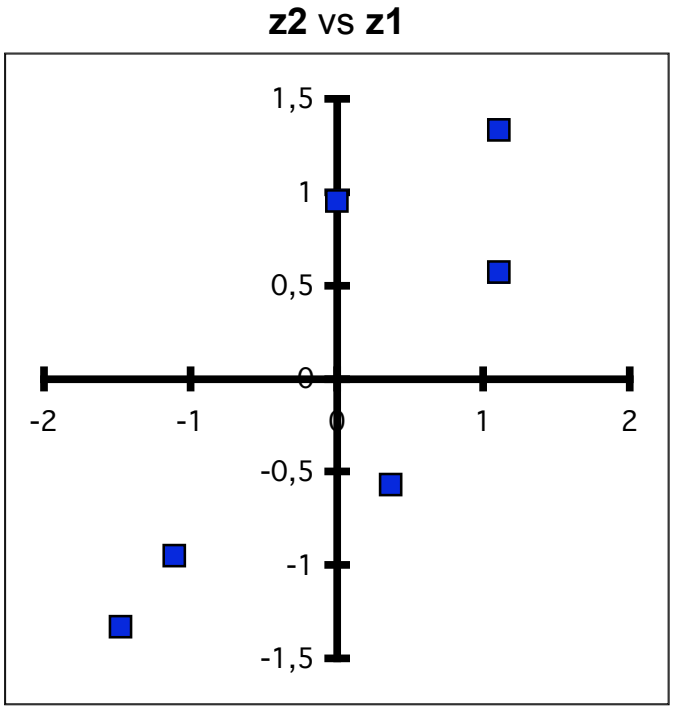
	Z		C			Y	
	0,369	-0,570				-0,142	-0,664
	1,108	0,570	0,71	-0,71		1,187	-0,380
	-1,477	-1,331	0,71	0,71	=	-1,985	0,103
	-1,108	-0,951				-1,456	0,111
	0,000	0,951				0,672	0,672
	1,108	1,331				1,724	0,158
moyenne	0,000	0,000				0,000	0,000
variance	1,000	1,000				1,819	0,181

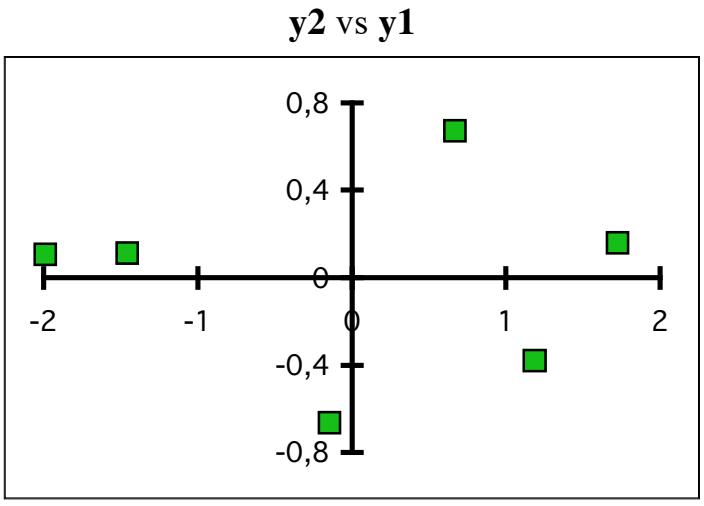
R_Z		R_Y	
1,000	0,819	1,000	0,000
0,819	1,000	0,000	1,000

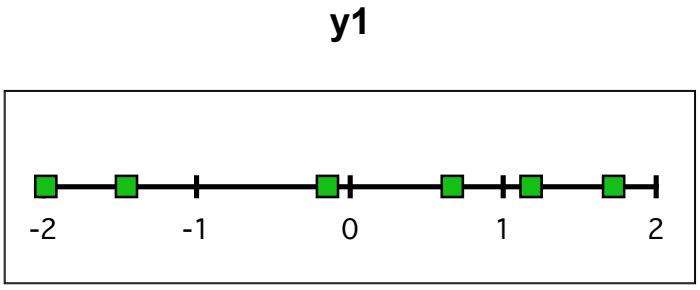
Nous constatons que les variables z1 et z2 sont corrélées (**R** = 0,819) et ont une variance 1, tandis que les variables y1 et y2 ne sont pas corrélées (**R** = 0) et ont une variance 1,819 et 0,181 respectivement. La variance totale (2) est inchangée.

La corrélation a donc été transformée en variance. Dans le système d'axe **Y** (composantes principales), on trouve une grande variance (beaucoup d'information) et une petite variance (bruit ?).

Graphiquement, on constate qu'après rotation, la plus grande variance se trouve sur l'axe **y**1, et que l'on peut pratiquement se passer de l'information reprise sur **y**2, de la même façon que pour identifier le chameau on peut se passer de le voir de dos lorsqu'on l'a vu de profil…







Les colonnes de **C** sont les vecteurs propres de la matrice de corrélation de **Z** associés aux valeurs propres qui sont les variances de **Y :**

Rz		c1		
1	0,819	0,71	=	1,29
0,819	1	0,71		1,29

c1		
0,71	1,819	= 1,29
0,71		1,29

Rz		c1		
1	0,819	-0,71	=	-0,13
0,819	1	0,71		0,13

c1		
-0,71	0,181	= -0,13
0,71		0,13

On peut donc considérer que les vecteurs propres dirigent une rotation orthogonale qui permet d'orienter un plan dans la direction de la plus grande variance.

La plus grande variance est la plus grande valeur propre de la matrice de corrélation. Elle est d'autant plus grande que la corrélation est forte.

La variance totale étant égale à 2 dans cet exemple, la variabilité de l'axe y1 représente 1,819/2 = 91% de la variabilité.

L'ACP peut se réaliser sur la matrice de variance covariance (données non réduites). Dans ce cas chacune des variables représente a priori un poids égal à sa propre variance, ce qui est généralement indésirable. Cette application est assez rare.

Plus de 2 variables

Nous avons abordé (module 220, diagonalisation) le problème de la recherche de vecteurs propres et valeurs propres qui représente déjà pour une matrice de corrélation 3 x 3 un problème non trivial. Nous laisserons dans une « boite noire » les problèmes théoriques et algorithmiques et considérons comme acquis qu'à partir d'une matrice de variance covariance (généralement calculée sur les données réduites : matrice de corrélation) il est possible de déterminer le jeu de valeurs propres – vecteurs propres avec la garantie que la première est la plus grande possible, la seconde la plus grande possible dans une direction orthogonale à la première, et ainsi de suite.

Notez toutefois que plusieurs algorithmes échouent lorsque la matrice R est singulière, ce qui sera le cas si le nombre d'observations est inférieur au nombre de variables, si une variable est une combinaison linéaire d'une autre, ou si deux variables ont une corrélation très proche de 1.

A partir d'une matrice de données **X** n x p (n observations pour p variables), généralement standardisée :

$$\underset{n\,x\,p}{\mathbf{Z}} = (\underset{n\,x\,p}{\mathbf{X}} - \underset{1\,x\,p}{\mathbf{m}x})/ \underset{1\,x\,p}{\mathbf{s}x}$$

On calcule la matrice de corrélation :

$$\underset{p\,x\,p}{\mathbf{R}} = 1/(n-1)\ \underset{p\,x\,n}{Z'}\ \underset{1\,x\,p}{\mathbf{Z}}$$

Et on la diagonalise :

$$\underset{p\,x\,p}{\mathbf{L}}\ \underset{p\,x\,p}{\mathbf{R}}\ \underset{p\,x\,p}{\mathbf{L'}} = \underset{p\,x\,p}{\mathbf{Q®}}$$

$$\underset{p\,x\,r}{\mathbf{F}} = \underset{p\,x\,r}{L\text{-}1}$$

r ≤ p représente le nombre de valeurs propres non nulles (rand de la matrice), ou le nombre de composantes principales que l'on décide arbitrairement de retenir.

Pour réaliser l'analyse, il est nécessaire de connaître

- la matrice de corrélation, qui expriment le degré de redondances entre les variables, permettant de réduire le nombre de dimensions du problème
- les valeurs propres : elles permettent de déterminer la proportion de variance exprimée dans un plan, ce qui est l'un des critères essentiel de sa représentativité
- la matrice orthogonale permettant de réaliser la rotation du système d'axes afin de placer un plan (généralement le plan formé des 2 premières composantes) dans la position adéquate.
- les coordonnées des observations (objets) projetées sur ce plan
- la relation entre ces nouveaux axes et les variables originales, afin de pouvoir les interprèter.

*Il est remarquable de constater que la matrice **F** contient toutes ces informations, « livrées en kit » et codées presque comme un génôme. Vous allez voir : c'est joli !*

F contient la matrice de corrélation :

$$\underset{p\,x\,p}{\mathbf{F}}\ \underset{p\,x\,p}{\mathbf{F'}} = \underset{p\,x\,p}{\mathbf{R}} \qquad \text{si } r < p \qquad \underset{p\,x\,r}{\mathbf{F}}\ \underset{r\,x\,p}{\mathbf{F'}} \approx \underset{p\,x\,p}{\mathbf{R}}$$

F contient la matrice des valeurs propres

$$\underset{r\,x\,p}{\mathbf{F'}}\ \underset{p\,x\,r}{\mathbf{F}} = \underset{r\,x\,r}{\Lambda}$$

F contient la matrice orthogonale de rotation des axes

$$\underset{p\,x\,r}{\mathbf{F}}\ \underset{r\,x\,r}{\Lambda^{-0,5}} = \underset{p\,x\,r}{\mathbf{C}} \text{ avec } \underset{p\,x\,r}{\mathbf{C}}\ \underset{r\,x\,p}{\mathbf{C'}} = \underset{p\,x\,p}{\mathbf{I}}$$

Celle-ci permet d'obtenir les coordonnées des objets après rotation :

$$\underset{n\,x\,p}{\mathbf{Z}}\ \underset{p\,x\,r}{\mathbf{C}} = \underset{n\,x\,r}{\mathbf{Y}}$$

Coordonnées standardisées (**m**y = à et **s**y = $\lambda^{0,5}$)

$$\underset{n\,x\,r}{\mathbf{Zy}} = (\underset{n\,x\,r}{\mathbf{y}} - \underset{1\,x\,r}{\mathbf{m}y})/\ \underset{1\,x\,r}{\mathbf{s}y}$$

F contient la matrice de corrélation entre les variables originales et les composantes principales

$$\underset{p\,x\,r}{\mathbf{F}} = 1/(n-1)\ \underset{p\,x\,n}{Z'}\ \underset{n\,x\,r}{\mathbf{Zy}}$$

Exemple de calcul

Soit une matrice de données X représentant le recensement de 15 recensements de 5 espèces (sur 5 sites, 3 recensements par sites).

	X	Esp 1	Esp 2	Esp 3	Esp 4	Esp 5
Site 1	recensement 1	7	5	2	3	51
	recensement 2	3	7	1	4	103
	recensement 3	2	18	4	3	65
Site 2	recensement 1	0	114	0	1	0
	recensement 2	0	145	0	0	0
	recensement 3	0	150	0	1	0
Site 3	recensement 1	0	88	2	27	1
	recensement 2	0	110	2	10	1
	recensement 3	0	111	6	59	4
Site 4	recensement 1	0	45	6	7	1
	recensement 2	0	11	4	2	3
	recensement 3	0	80	2	3	3
Site 5	recensement 1	8	41	0	4	3
	recensement 2	19	70	0	3	7
	recensement 3	6	58	0	3	8
	Moyenne	3	70,2	1,933	8,667	16,67
	variance	27,71	2404	4,638	236,8	959,1

Le tableau est standardisé pour donner à chaque espèce le même poids dans la variance totale.

	Z	Esp 1	Esp 2	Esp 3	Esp 4	Esp 5
Site 1	recensement 1	0,8	-1,3	0	-0,4	1,1
	recensement 2	0	-1,3	-0,4	-0,3	2,8
	recensement 3	-0,2	-1,1	1	-0,4	1,6
Site 2	recensement 1	-0,6	0,9	-0,9	-0,5	-0,5
	recensement 2	-0,6	1,5	-0,9	-0,6	-0,5
	recensement 3	-0,6	1,6	-0,9	-0,5	-0,5
Site 3	recensement 1	-0,6	0,4	0	1,2	-0,5
	recensement 2	-0,6	0,8	0	0,1	-0,5
	recensement 3	-0,6	0,8	1,9	3,3	-0,4
Site 4	recensement 1	-0,6	-0,5	1,9	-0,1	-0,5
	recensement 2	-0,6	-1,2	1	-0,4	-0,4
	recensement 3	-0,6	0,2	0	-0,4	-0,4
Site 5	recensement 1	0,9	-0,6	-0,9	-0,3	-0,4
	recensement 2	3	-0	-0,9	-0,4	-0,3
	recensement 3	0,6	-0,2	-0,9	-0,4	-0,3
	Moyenne	0	0	0	0	0
	variance	1	1	1	1	1

La matrice de corrélation peut être calculée par la fonction Excel coefficient.correlation( ), mais cela peut être fastidieux. Le calcul matriciel est plus direct : R=PRODUITMAT(TRANSPOSE(Z);Z)/n

1	-0,3	-0,4	-0,2	0,1
-0,3	1	-0,3	0,2	-0,6
-0,4	-0,3	1	0,5	0
-0,2	0,2	0,5	1	-0,2
0,1	-0,6	0	-0,2	1

La matrice **F** doit être obtenue par un logiciel ad hoc. Il n'est cependant pas toujours aisé de savoir quels vecteurs propres sont délivrés par le listing. Pour le contrôler, utiliser la propriété :

$$\mathbf{F'\ F = \Lambda}$$

Il s'avère souvent utile de maîtriser ses propres résultats dans le tableur plutôt que de dépendre des outputs du logiciel.

F

0,6	0,2	0,7	0	-0,1
-0,7	0,7	-0,1	-0,3	-0,2
-0,5	-0,8	0,1	0,3	-0,2
-0,7	-0,4	0,5	-0,3	0,1
0,6	-0,6	-0,2	-0,5	-0,1

Λ

1,92	0	0	0	0
0	1,64	0	0	0
0	0	0,81	0	0
0	0	0	0,51	0
0	0	0	0	0,13

On peut vérifier que F représente les vecteurs propres de R par la relation :

$$\mathbf{R\ f}_i, = \mathbf{f}_i, \lambda_i$$

La matrice **C** est obtenue en normant les vecteurs propres de **F**., ce qui revient à diviser par la racine carrée de leur norme λ. A cause des valeurs nulles, on en peut pas calculer directement $\Lambda^{-0,5}$ Le plus simple techniquement est de créer un vecteur λ' 1 x r en faisant la $^{-0,5}$des lignes de Λ et de faire l'opération entre scalaires **F**/ $\lambda'^{-0,5}$.

C

0,5	0,2	0,8	0	-0,3
-0,5	0,5	-0,1	-0,4	-0,6
-0,3	-0,6	0,1	0,4	-0,6
-0,5	-0,3	0,5	-0,5	0,4
0,5	-0,4	-0,2	-0,7	-0,3

On peut vérifier qu'il s'agit d'une matrice orthogonale par la relation :

C **C'** **= I**
p x r r x p p x p

Eventuellement, ces opérations peuvent être réalisées sur un petit nombre de colonnes, voire seulement 2. Dans ce cas, le produit

F **F'** **≈ R**
p x r r x p p x p

La différence avec la matrice **R** originale permet de déterminer dans quelle mesure les variables ont été préservées dans l'espace réduit.

Il est maintenant possible de calculer les coordonnées des objets :

Z C = Y
n x p p x r n x r

	Y	composante principale 1	2	3	4	5
Site 1	recensement 1	1,658	-0,950	0,233	-0,023	0,064
	recensement 2	2,168	-1,533	-0,796	-1,422	0,061
	recensement 3	0,993	-1,771	-0,580	-0,082	-0,442
Site 2	recensement 1	-0,388	1,327	-0,724	-0,127	0,124
	recensement 2	-0,658	1,676	-0,797	-0,331	-0,278
	recensement 3	-0,739	1,709	-0,769	-0,401	-0,315
Site 3	recensement 1	-1,267	-0,084	0,274	-0,388	0,543
	recensement 2	-0,932	0,497	-0,338	-0,018	-0,136
	recensement 3	-3,093	-1,714	1,479	-0,869	-0,028
Site 4	recensement 1	-0,814	-1,311	-0,201	1,342	-0,447
	recensement 2	0,013	-1,009	-0,425	1,330	0,344
	recensement 3	-0,384	0,293	-0,556	0,389	0,042
Site 5	recensement 1	0,969	0,721	0,673	0,335	0,571
	recensement 2	1,740	1,367	2,256	0,151	-0,518
	recensement 3	0,734	0,782	0,272	0,112	0,413
	Moyenne	0	0	0	0	0
	variance	1,923	1,635	0,807	0,506	0,128
	% cumulés	38%	71%	87%	97%	100%

Constatons que les variances de **Y** sont bien les valeurs propres de **R**.
Il est fréquent de les exprimer en % cumulés de façon à indiquer l'information récupérée dans un espace réduit à 1, 2, 3 … dimensions.

Plusieurs logiciels présentent, parfois exclusivement, les composantes principales réduites **Y**z. Il suffit de diviser **Y** par $\lambda'^{-0,5}$ comme déjà fait pour passer de **F** à **C.**

Cette transformation ne se justifie guère que pour faciliter les graphiques dans d'anciennes procédures. Par contre ce calcul permet de vérifier par calcul que **F** représente bien les corrélations entre **Z** et **Y**. Cette information est très importante dans l'interprétation des composantes principales. En effet ces dimensions « fictives » prendront une valeur concrète au vu de leurs corrélation avec les variables originales.

Interprétation

Trois éléments d'interprétation se complètent pour commenter la représentation des objets dans un espace réduit :

- **la variabilité reconstituée :** valeurs propres en pourcentage cumulé. Aucune règle n'existe en la matière. Parfois, une grande variance (90% par exemple) peut être associée à une information triviale, comme la taille des individus (voir l'exemple sur l'analyse des mensurations de mâchoires de singes) et l'information pertinente se trouver sur la 2° voire la 3° composante. Si le nombre de variable est grand, le % de variabilité reconstitué par une composante peut être faible. Sur 100 variables standardisées, chacune n'apporte qu'un % d'information : deux composantes apportant ensemble 15% d'information peuvent être intéressantes…
- **la représentation des objets (observations)** dans un espace réduit. Les objets sont projetés sur le plan Y1 vs Y2, parfois aussi Y1 vs Y3 et Y2 vs Y3. Il est rare que la quatrième composante soit encore informative. L'information est généralement graphique. La proximité des points est le signe que les objets représentés ont le même profil de mesures, pour les variables bien corrélées au plan. Il ne faut toutefois pas perdre de vue que des objets proches sur le plan peuvent être éloignés suivant d'autres dimensions. Le regroupement d'objets formant des clusters sur le plan est souvent une base d'interprétation intéressante. Cette analyse est souvent graphique, bien que de nombreux logiciels ne facilitent pas cette approche.
- **La corrélation entre les variables et les composantes.** Lorsque qu'une variable est fortement corrélée à un axe (composante) on peut associer l'axe et la variable. On interprétera ainsi la répartition des objets suivant une composante comme leur répartition suivant les variables originales fortement corrélées à l'axe. Cette lecture facilite l'interprétation de la répartition des objets. Certains logiciels proposent d'ailleurs une représentation simultanée des objets et des variables sur le même graphique. Comme pour les objets, la proximité des points-variables indique une forte corrélation entre ces variables. On peut alors considérer qu'il s'agit d'une représentation graphique de la matrice de corrélation.

Limites de l'interprétation

Les deux propriétés :

$$\mathbf{F}\ \mathbf{F'}\ = \mathbf{R}$$
$$\mathbf{F'}\ \mathbf{F}\ = \Lambda$$

Ont le corollaire suivant :

F	Y1	Y2	Y3	Y4	Y5	Σf_j^2
Z1	0,6	0,2	0,7	0	-0,1	1
Z2	-0,7	0,7	-0,1	-0,3	-0,2	1
Z3	-0,5	-0,8	0,1	0,3	-0,2	1
Z4	-0,7	-0,4	0,5	-0,3	0,1	1
Z5	0,6	-0,6	-0,2	-0,5	-0,1	1
Σf_i^2	1,923	1,635	0,807	0,506	0,13	5

Ceci implique ,

- suivant les colonnes de F qu'une composante principale de faible valeur propre ne peut pas être fortement corrélée aux variables, ce qui limite sérieusement son intérêt.
- suivant les colonnes de F , qu'une variable qui est déjà fortement corrélée à la 2°, 3° composante ne peut plus l'être aux autres (puisque la somme des carrés est imitée à 1) ou réciproquement que si elle ne s'est pas « exprimée » sur ces plans elle s'exprime forcément ailleurs.

On peut d'ailleurs illustrer la reconstitution des variables dans un espace réduit à partir d'une matrice **F** p x r :

F	Y1	Y2	Σf_j^2
Z1	0,64	0,23	0,46
Z2	-0,66	0,67	0,88
Z3	-0,46	-0,81	0,87
Z4	-0,69	-0,40	0,64
Z5	0,63	-0,57	0,71
Σf_i^2	1,92	1,64	3,56

Ceci montre par exemple que la variable originale N°1 (espèce 1) est moins bien reconstituée que les autres.

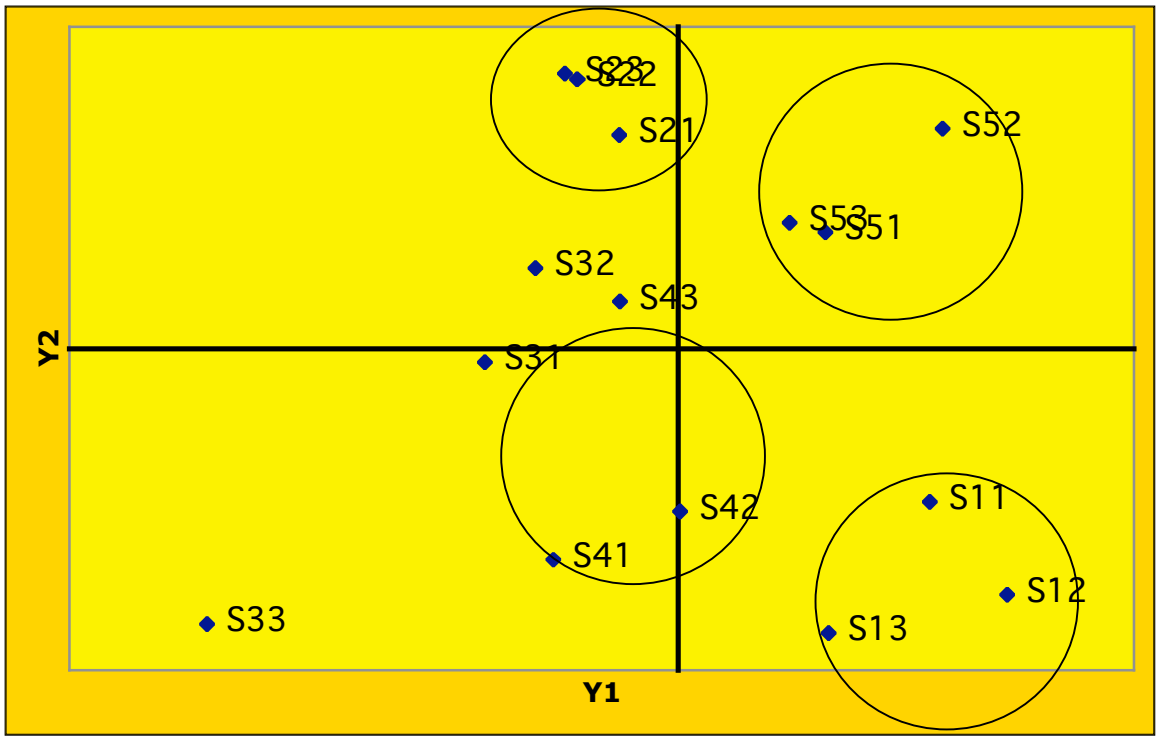
L'ACP reste une technique assez intuitive. Il en va des plans factorielles comme des coupes au microscope : le bon angle et la bonne coloration ne sont pas découverts du premier coup. Il ne faut donc pas s'attendre à ce que la première analyse livre le bon résultat. Il faut être attentifs aux statistiques uni- et bivariées des données originales. Des objets méritent d'être enlevés, traités par sous groupes ; des variables transformées, voire supprimées . Ce n'est donc pas UNE ACP qui sera réalisée, mais toute une série. D'où la nécessité de disposer d'un bon logiciel, spécialement sur le plan graphique. Ce n'est pas évident.

Analyse graphique des résultats.

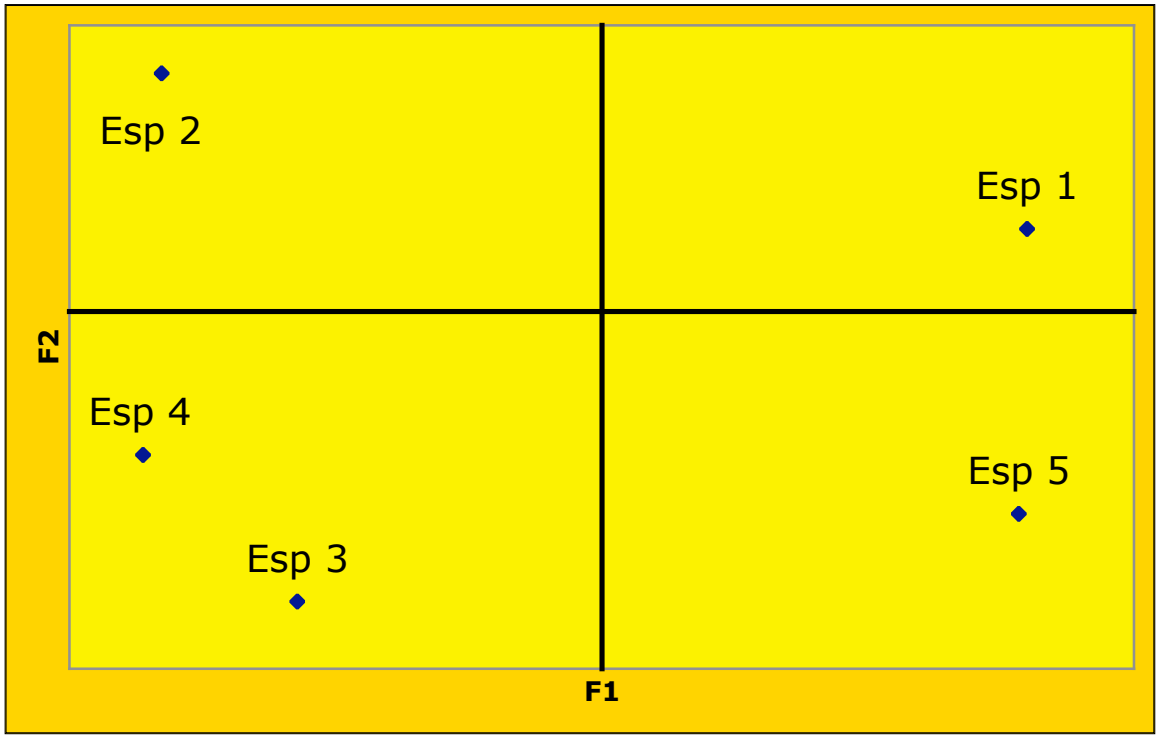
Les matrices de données soumises à l'analyse sont par définition de grandes, parfois très grandes matrices. Des centaines d'objets, des dizaines de variables sont un ordre de grandeur fréquent. Il est donc indispensable de pouvoir imprimer un LABEL sur les points graphiques, ce qui est généralement le facteur limitant des logiciels.

Sur le tableur Excel, cette fonction n'existe pas d'origine. Le site http://www.appspro.com/Utilities/ChartLabeler.htm propose un plugin pour Mac et PC, relativement simple à utiliser.

Représentation graphique des objets : Le premier chiffre du label représente le site et le second l'un des 3 recensements.



Le graphique indique clairement que les recensements d'un même site se ressemblent. Le site 3 est plus hétérogène.



Le nombre de variables ne justifie pas vraiment le recours à un graphique. Cependant celui-ci met en évidence que l'axe 1 indique globalement, de gauche à droite, une diminution des espèces 2,3,4 et l'augmentation des espèces 1 et 5 (attention, la matrice F indique que l'espèce 1 est relativement mal reconstituée sur ce plan).

L'axe 2 indique, du bas vers le haut, une diminution des espèces 3, 4, 5 et une augmentation des espèces 1 et 2.

Ces variations d'abondance sont à mettre en relation avec la répartition gauche/droite, haut/bas des points stations sur le graphique Y.

Notez toutefois que cet exemple limité a été choisi pour illustrer le calcul sur de petites matrices.

Des situations permettant une interprétation plus détaillée sont présentées sur les feuilles d'exercices.