Analyse en composantes principales
(TP Splus)

Analyse en composantes principales des données Forsythia

54 variétés de Forsythia ont été mesurées pour 9 caractères morphologiques floraux :

- LP = longueur du pédoncule floral
- Lcal = longueur du calice
- lsep = largeur des sépales
- LTC = longueur totale de la corolle (tube + limbe)
- Pgl = plus grande largeur du limbe
- LTP = longueur du pistil (le forsythia est hétérostylé avec des formes brevi ou longistylées)
- LTE = longueur totale de l'étamine
- LF = longueur du filet
- Del = différence de hauteur entre le stigmate et les anthères. Le signe négatif indique que la variété est brévistylée (stigmate sous les anthères)

Le but de cette analyse est d'essayer d'étudier la structuration de la diversité des différentes variétés étudiées.

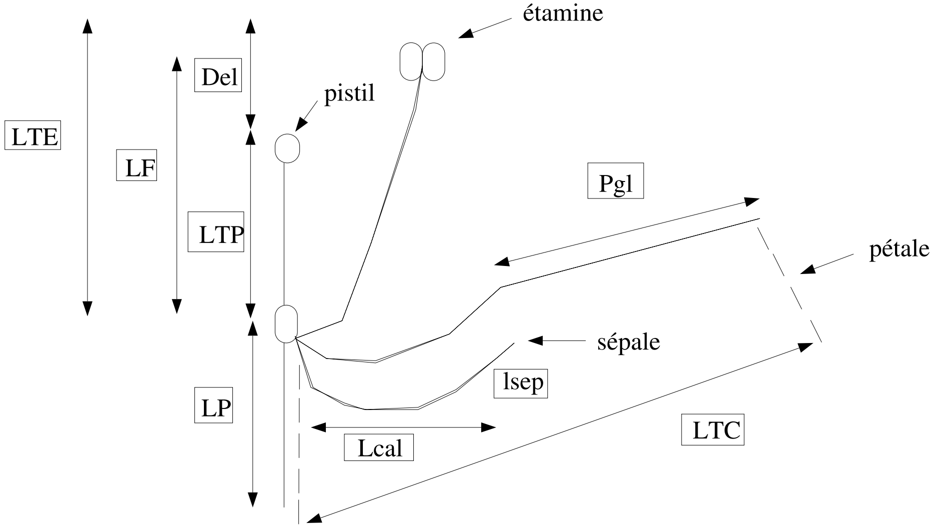
Sous Unix :

Unix> vi forsy (visualisation du fichier de données forsy avec l'éditeur vi)

```
       LP Lcal lsep   LTC   PGl  LTP  LTE   LF   Del
 C18  6.30 4.45 1.75 21.10  6.50 3.80 5.70 4.10 -3.15
 C17  6.95 6.75 2.53 18.30  7.95 3.30 6.70 4.30 -4.20
 C16  7.35 6.70 2.42 16.95  7.15 3.30 7.00 4.70 -4.45
 C27  8.55 6.20 3.26 26.35  9.30 4.40 6.30 4.25 -3.45
.....................................................
.....................................................
 C51 10.10 7.45 3.01 25.55 11.25 8.65 4.20 1.55  2.95
 C52 12.80 8.05 2.42 17.55  7.70 4.40 6.80 4.20 -4.25
 C53  7.65 5.10 2.57 22.10  6.55 8.00 3.45 1.45  3.55
 C54  4.40 4.50 2.75 13.45  5.10 6.50 2.75 1.15  2.55
 C56 11.50 7.60 3.87 29.90 12.60 5.40 5.05 2.00 -1.20
```


- Faire le schéma au tableau.

Schéma de la fleur de Forsythia



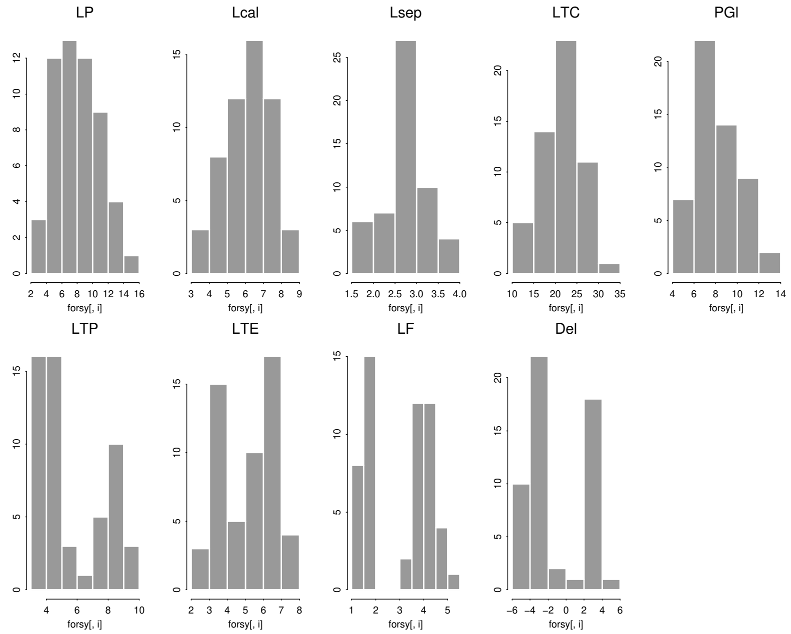
Passer sous Splus :

Récupération du fichier de données sous forme de dataframe

```
>forsy_read.table("forsy",header =T) #on rapatrie le fichier
 et les noms de lignes et de colonnes ; on l'appelle forsy

>forsy[1:10,] #visualisation des 10 premières lignes du fichier
               de données forsy
```

Représentation graphique des données de forsy

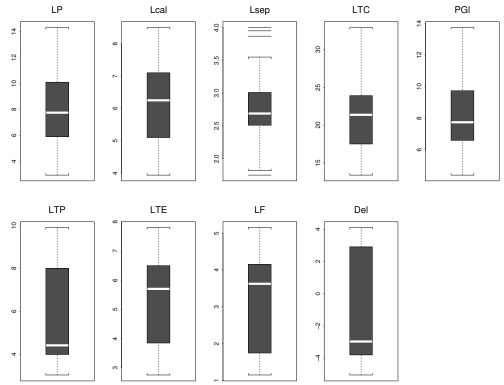
```
>motif() #ouverture d'une fenêtre graphique
>par(mfrow=c(2,5)) #préparation de la fenêtre graphique :
                    division en 10 (2 x 5)
>for (i in 1:9){hist(forsy[,i],main = names(forsy[i]))} #boucle
 pour sortir les histogrammes de toutes les variables
```



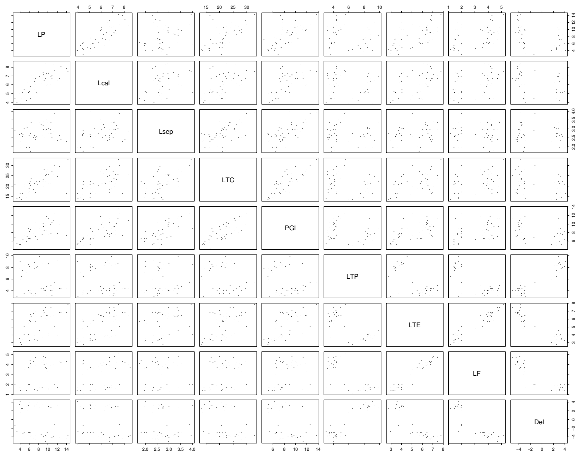
On observe que les 5 variables LP, Lcal,Lsep, LTC, Pgl suivent une distribution de type normale. Nous n'observons pas de forte asymétrie; il n'est donc pas nécessaire de faire des transformations (type log ou racine). En effet, bien qu'il n'y est pas d'hypothèses particulières sur la distribution des variables pour réaliser une ACP, la présence de variables à distribution asymétrique peut biaiser les corrélations.

Les 4 autres variables par contre sont nettement bimodales. Ces 4 variables concernent les mesures de longueurs relatives aux étamines ou au pistil. Les 2 groupes mis en évidence dans ces distributions correspondent aux types brévistylés (étamines plus courtes que pistil) et longistylés (étamines plus longues que le pistil).

```
> par(mfrow=c(2,5))
> for (i in 1:9){boxplot(forsy[,i],main=names(forsy[i]))}
  #représentation en boxplot
```




```
> pairs(forsy)#représentation des biplots 2 à 2
```



Autre forme de représentation qui renseigne toujours sur la symétrie des distributions mais non sur l'uni/bimodalité.

Avec la fonction pairs on remarque :

- les fortes corrélations existant entre certaines variables : (LP,Lcal), (LTC,PGl), (LTE,LF).
- les distributions bimodales des variables vues ci-dessus. On voit également sur les graphiques en bas à droite comment la bimodalité peut influer sur les corrélations entre variables ; on peut envisager également de demander au stagiaire la matrice de corrélation entre variables :

```
> forsymat_as.matrix(forsy) #transformation du dataframe en
  matrice
> round(cor(forsymat),2) #matrice de corrélation avec arrondi
  à 2 chiffres
```

	LP	Lcal	lsep	LTC	PGl	LTP	LTE	LF	Del
LP	1.00	0.82	0.34	0.65	0.67	-0.09	0.58	0.45	-0.34
Lcal	0.82	1.00	0.45	0.56	0.66	-0.10	0.57	0.42	-0.34
lsep	0.34	0.45	1.00	0.49	0.69	0.04	0.28	0.14	-0.11
LTC	0.65	0.56	0.49	1.00	0.79	-0.17	0.54	0.41	-0.36
PGl	0.67	0.66	0.69	0.79	1.00	-0.22	0.65	0.52	-0.45
LTP	-0.09	-0.10	0.04	-0.17	-0.22	1.00	-0.69	-0.78	0.94
LTE	0.58	0.57	0.28	0.54	0.65	-0.69	1.00	0.96	-0.88
LF	0.45	0.42	0.14	0.41	0.52	-0.78	0.96	1.00	-0.91
Del	-0.34	-0.34	-0.11	-0.36	-0.45	0.94	-0.88	-0.91	1.00

Sur cette matrice de corrélation, identifier les variables qui présentent une liaison linéaire positive forte (ex : LTE et LF ; Del et LTP) ou une forte liaison linéaire négative (ex : Del et LF).

Séparation des variétés brévi et longistylées

```
> forsyor_forsy[order(forsy$Del),] #on ordonne forsy par la
                                    variable Del
> forsyor[1:8,]
       LP Lcal lsep   LTC   PGl  LTP  LTE   LF   Del
C11  9.50 6.70 2.99 25.10 11.35 4.00 7.45 4.70 -5.05
C37  9.20 6.45 2.55 28.05  8.50 4.05 6.80 4.00 -4.50
C16  7.35 6.70 2.42 16.95  7.15 3.30 7.00 4.70 -4.45
C52 12.80 8.05 2.42 17.55  7.70 4.40 6.80 4.20 -4.25
C17  6.95 6.75 2.53 18.30  7.95 3.30 6.70 4.30 -4.20
C10 14.00 7.85 2.30 16.45  7.35 4.45 6.85 4.30 -4.20
 C4 10.90 6.70 3.00 22.55 10.35 4.05 6.40 3.95 -4.10
C35  9.10 7.10 2.97 22.45  9.45 4.30 6.75 4.45 -4.00
> forsyor[50:54,]
      LP Lcal lsep   LTC  PGl  LTP  LTE   LF  Del
C53 7.65 5.10 2.57 22.10 6.55 8.00 3.45 1.45 3.55
C23 8.75 6.75 2.67 22.50 7.55 8.75 3.70 1.70 3.60
C15 7.85 5.05 2.54 20.75 6.35 8.20 3.30 1.40 4.00
C48 6.05 5.25 2.00 17.55 8.25 8.20 3.30 2.00 4.00
C33 4.05 3.95 2.00 16.35 6.25 9.05 3.60 1.75 4.10
> forbrev_forsyor[forsyor$Del<0,] #création du fichier des brevi
> forlong_forsyor[forsyor$Del>0,] #création du fichier des longi
> forbrev_forbrev[,-9] #élimination de la variable Del
> forlong_forlong[,-9] #élimination de la variable Del
```

ACP sur les brévistylés

```
> library(multidim,first=T) #on appelle la librairie multidim
> search() #on verifie la position de multidim
[1] ".Data"
[2] "/bao/prive/splus3.2/library/multidim/.Data"
[3] "/bao/prive/splus3.2/splus/.Functions"
[4] "/bao/prive/splus3.2/stat/.Functions"
[5] "/bao/prive/splus3.2/s/.Functions"
[6] "/bao/prive/splus3.2/s/.Datasets"
[7] "/bao/prive/splus3.2/stat/.Datasets"
[8] "/bao/prive/splus3.2/splus/.Datasets"
```


L'analyse des représentations graphiques précédentes nous conduit à identifier 2 groupes d'individus : le groupe des brévistylés et celui des longistylés.

Compte tenu de la très forte structuration du fichier, l'ACP sur toutes les données ne fait que traduire le clivage vu précédemment entre brévi et longistylés (il est possible de faire réaliser cette ACP globale par les stagiaires; mais on ne fait que reproduire ce que l'on observe dans les graphiques précédents). Plusieurs stratégies d'analyse peuvent être envisagées :

- travailler sur l'ensemble des données en éliminant les variables relatives aux longueurs des étamines et du pistil de manière à ne plus faire apparaître la structuration;
- séparer les variétés en 2 groupes; et effectuer une ACP sur chacun des groupes ; c'est cette voie que nous avons choisi. La séparation des 2 groupes se fait à l'aide de la variable Del (<0 pour les brévi; >0 pour les longi).

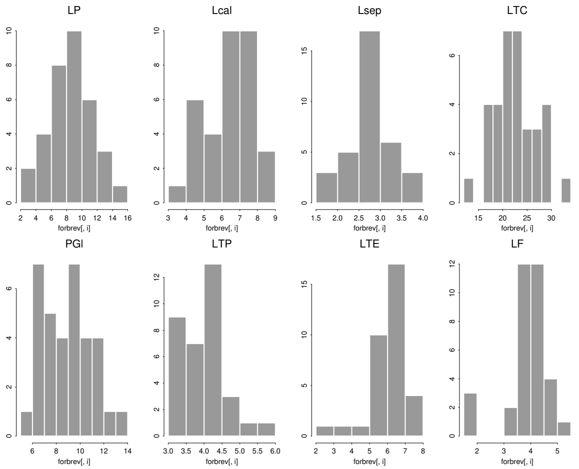
On élimine la variable Del dans les 2 fichiers car construite à partir de variables déja présentes et en plus très corrélée avec.

On effectue ensuite les ACP. Noter que le fait de diviser en 2 le fichier va diminuer le nombre d'individus par analyse; mais dans les 2 cas le nombre d'individus reste toujours supérieur à celui des variables; Dans le cas de cette analyse, il n'est pas souhaité de donner plus de poids à une variable plutôt qu'à une autre; nous travaillerons donc sur les variables centrées réduites (option proposée par défaut dans la fonction ACP).

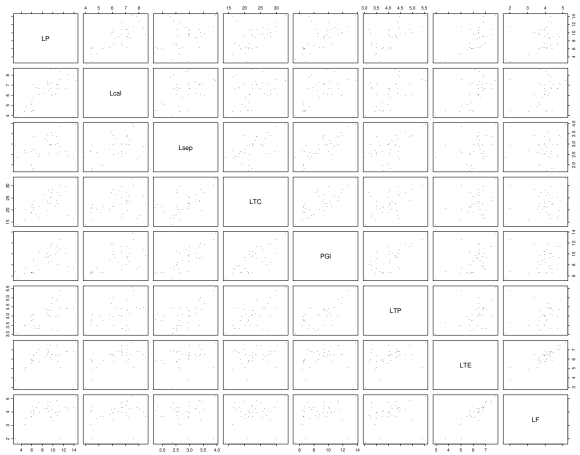
Avec search, on vérifie la position de la librairie multidim. On a fait first=T car certaines fonctions de multidim existent déjà sous le même nom dans Splus.

Représentations graphiques préliminaires sur le groupe des brévistylés

```
> par(mfrow=c(2,4)) #on recrée une fenêtre adaptée
> for(i in 1:8){hist(forbrev[,i],main=names(forbrev[i]))}
  #histogrammes des 8 variables
```

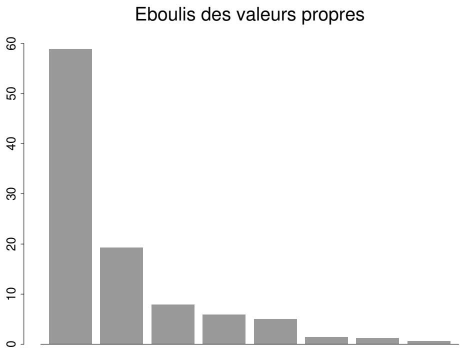



```
> pairs(forbrev)
```



ACP sur le groupe des brévistylés

```
> forbrev.acp_acp(forbrev) #lancement de l'ACP sur les brévi
Contributions des lignes (en millièmes)

C11 C37 C16 C52 C17 C10 C4 C35 C5 C44 C28 C13 C24 C40 C3 C47  C2 C20 C14 C6
 18   9  25  27  15  36  5   6 22   3  17  20   5  12 32  13 119   8   7 17
C39 C27 C50 C18 C8 C29 C45 C36 C32 C55 C25 C41 C38 C56
 11   8  46  34 22  25  22  31  25  31  24 135  76  92

Contributions des colonnes (en millièmes)
  LP Lcal lsep LTC PGl LTP LTE  LF
 125  125  125 125 125 125 125 125

1: On continue ?
2: Arret
Selection: 1
Inertie totale:  8
> prexpl(forbrev.acp)
Pourcentage d'inertie expliquée
            f1 f2 f3 f4 f5 tot
inertie exp 59 19  8  6  5  97
inertie cum 59 78 86 92 97  97
> par(mfrow=c(1,1)) #on recadre la fenêtre graphique sur
                         toute la page
> eboulis(forbrev.acp) #éboulis des valeurs propres
                            correspondant aux axes
```



La fonction ACP() travaille par défaut sur les variables centrées réduites. Si l'on ne réduit pas les variables, il faut faire acp(forbrev,reduc=F).

La fonction acp() donne des renseignements sur la contribution des individus (des lignes) d'une part et des variables (des colonnes) d'autre part. Sur les contributions des lignes, on remarque que 4 individus (C2, C41, C38 et C56)représentent 42 % de ce que va décrire l'analyse.

En cas de contributions trop déséquilibrées, le programme propose d'interrompre l'analyse.

La fonction prexpl() donne le pourcentage de variance expliquée par les différentes composantes principales et les pourcentages cumulés.

La visualisation graphique proposée par la fonction eboulis() est plus parlante.

Dans le cas de notre étude les 2 premiers axes expliquent à eux seuls 78 % de la variabilité; nous pouvons donc nous contenter de la représentation sur le premier plan 1-2.

```
> contri(forbrev.acp)

Cosinus carré individus axes principaux

     f1  f2  f3  f4  f5 Axes >5  TOT
C11 608 110  91  85  34      72 1000
C37 125  42  79 160 426     168 1000
C16 151 724  11 110   0       5 1000
C52  72 359 384  91  91       4 1000
C17 223 510  29 206   8      25 1000
C10  48 385 314 132 102      19 1000
..........
..........

Contribution des variables aux facteurs

       f1   f2   f3   f4   f5  f>5  TOT VARI
  LP  159   14  146  169    3  229  125  232
Lcal  128   20  499   25   20  133  125   44
lsep  104  182    1  395   55  100  125    8
 LTC  115   89  123   62  508   42  125  543
 PGl  170   67   14   39    1  263  125  119
 LTP  131   42   30  262  410   55  125   10
 LTE  129  215   37   19    2  109  125   27
  LF   64  372  149   29    1   70  125   16
     1000 1000 1000 1000 1000 1000 1000 1000
```


La fonction contri() donne les résultats numériques de l'ACP ($\cos^2$, contributions des individus; correlations^2, contributions des variables).

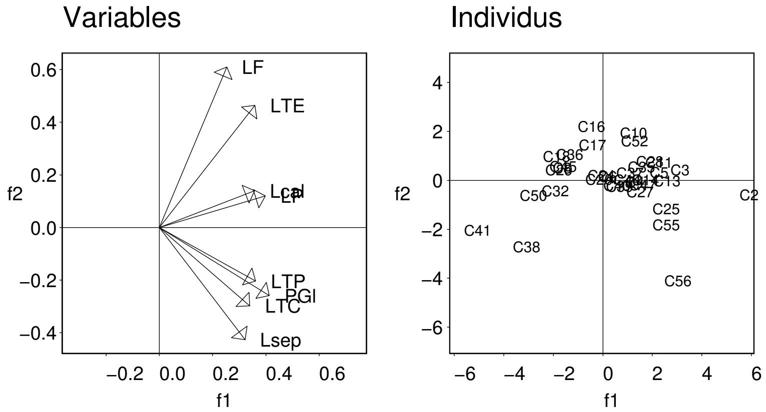
Distribuer aux stagiaires les sorties papier de l'analyse (Annexe TP).

On peut analyser à ce stade la contribution des individus et colonnes ou le faire en parallèle de la représentation graphique.

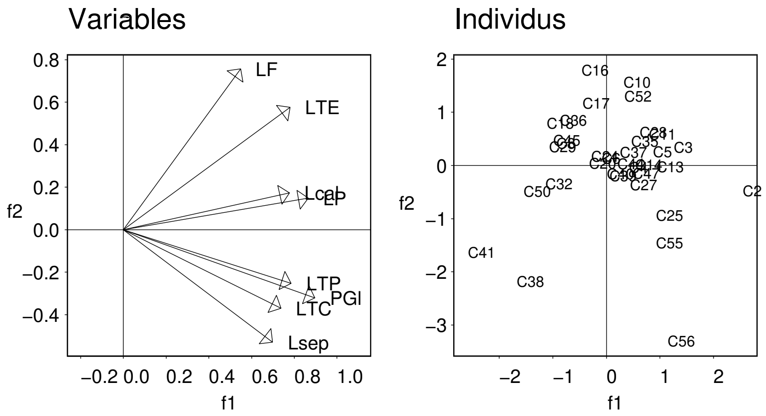
Pour l'axe 1, la somme des contributions de C2, C41, C38 est égale à 47 %. Pour l'axe 2, la somme des contributions de C41, C38, C56 est égale à 54 % (avec une contribution de C56 égale à 32 %).

Quatre variétés jouent un rôle important : C2 est une grosse fleur, C38, C41 et C56 sont les seules fleurs de l'étude à avoir un pistil de la taille des étamines ; elles possèdent des étamines dont les proportions sont proches de celles des longistylées.

```
> biplot(forbrev.acp) #on choisit un biplot isométrique ligne

1: Biplot isometrique ligne ?
2: Biplot isometrique colonne ?
3: Biplot a_l'hollandaise ?
Selection: 1
```




```
> biplot(forbrev.acp)

1: Biplot isometrique ligne ?
2: Biplot isometrique colonne ?
3: Biplot a_l'hollandaise ?
Selection: 2
```



On se propose d'analyser les représentations graphiques et comparer les sorties biplot isométriques ligne et colonne.
On reprécisera aux stagiaires le fait que dans le biplot isométrique ligne, on a une représentation exacte des individus ; dans le biplot isométrique colonne, une représentation exacte des variables (+ présence du cercle des corrélations). Globalement, il y a peu de différence de représentations.

Informations sur les variables :

- bonne représentativité sur les 2 axes (cf données de contri + biplot colonne)
- contribution des variables aux 2 axes (cf données de contri + biplot colonne)
- corrélations entre variables : on peut redemander aux stagiaires de sortir la matrice de corrélation des variables sur le fichier forbrev :

```
> round(cor(as.matrix(forbrev)),2)

       LP Lcal lsep  LTC  PGl  LTP  LTE   LF
  LP 1.00 0.80 0.42 0.55 0.66 0.66 0.68 0.47
Lcal 0.80 1.00 0.46 0.39 0.60 0.44 0.63 0.40
lsep 0.42 0.46 1.00 0.58 0.82 0.59 0.29 0.05
 LTC 0.55 0.39 0.58 1.00 0.77 0.60 0.38 0.18
 PGl 0.66 0.60 0.82 0.77 1.00 0.73 0.53 0.29
 LTP 0.66 0.44 0.59 0.60 0.73 1.00 0.47 0.25
 LTE 0.68 0.63 0.29 0.38 0.53 0.47 1.00 0.90
  LF 0.47 0.40 0.05 0.18 0.29 0.25 0.90 1.00
```

  on s'aperçoit que les variables très corrélées à partir des données brutes le sont aussi dans le résultat de l'ACP ((LP,Lcal), ...) et inversement ((LF,LTP), ...).

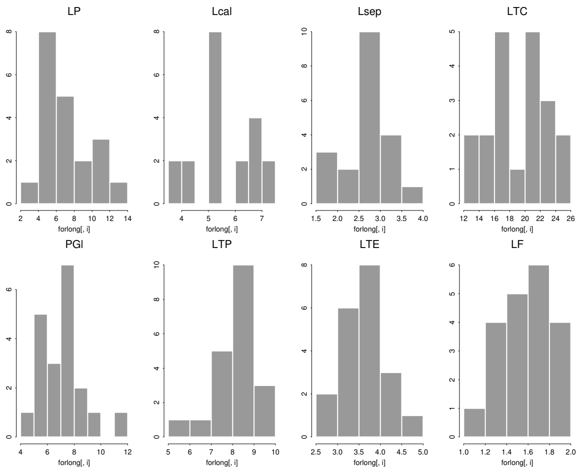
- le facteur 1 est un facteur taille : toutes les variables sont orientées dans le même sens.
- le facteur 2 est apparemment un facteur forme qui sépare d'une part LF, LTE, Lcal, LP et d'autre part LTP, PGl, LTC, lsep. Toutefois, il est peut être influencé par des individus atypiques (C41, C38, C56).

Informations sur les individus (à analyser avec les résultats de contri) :
- bonne répartition du nuage; beaucoup de points au centre.

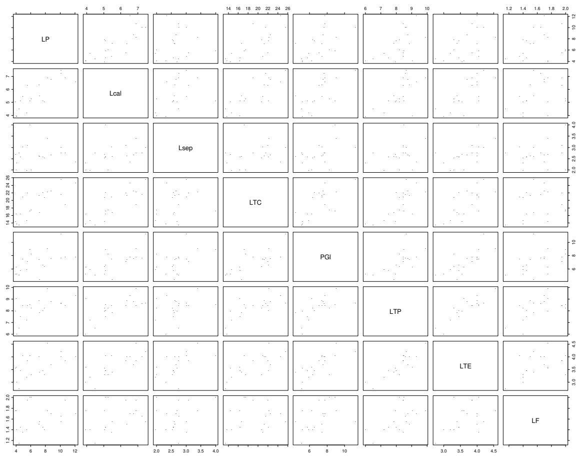
Remarquons que la forme du nuage (type fer à cheval) semble faire ressortir une non linéarité des liaisons.

Représentations graphiques préliminaires sur le groupe des longistylés

```
> par(mfrow=c(2,4))
> for (i in 1:8){hist(forlong[,i],main=names(forlong[i]))}
```




```
> pairs(forlong)
```



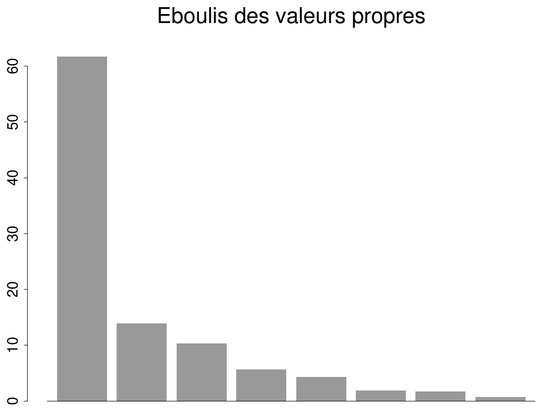
Présenter rapidement les graphiques.

Attention au plus faible nombre d'individus.

On retrouve les variables corrélées.

ACP sur les longistylés

```
> forlong.acp_acp(forlong) #lancement de l'ACP sur les longistyl
Contributions des lignes (en millièmes)

C26 C21 C30 C54 C22 C42 C34 C51 C7 C43 C49 C9 C1 C15p C31 C53 C23 C15 C48 C33
111  27  25 116  24  40  91 108 34  14  62 85 66   25  34  13  20  16  41  51

Contributions des colonnes (en millièmes)

  LP Lcal lsep LTC PGl LTP LTE  LF
 125  125  125 125 125 125 125 125

1: On continue ?
2: Arret
Selection: 1
Inertie totale:  8
> prexpl(forlong.acp)
Pourcentage d'inertie expliquée

            f1 f2 f3 f4 f5 tot
inertie exp 62 14 10  6  4  96
inertie cum 62 76 86 92 96  96
> par(mfrow=c(1,1))
> eboulis(forlong.acp)
```

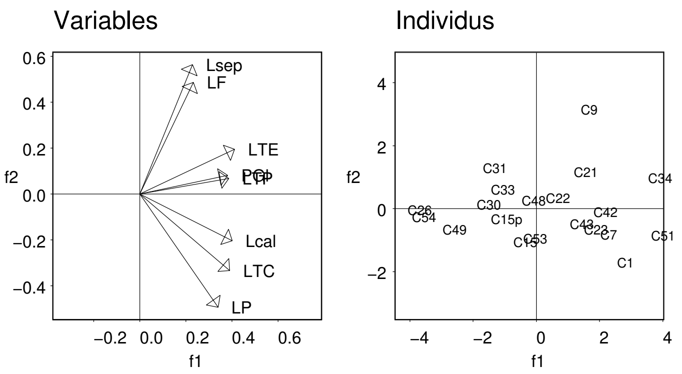


La cassure se fait réellement entre le $1er$ et le $2ième$ axe. Il y a peu de différence entre les axes 2 et 3. On choisit donc 3 axes.

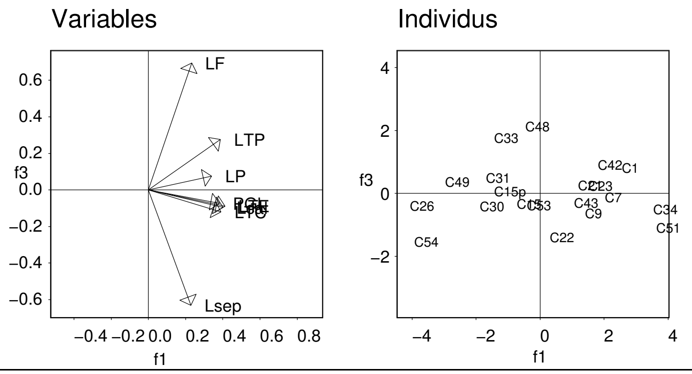
```
> contri(forlong.acp)

 Cosinus carré individus axes principaux

      f1  f2  f3  f4  f5 Axes >5  TOT
 C26 937   0   9  22  15      18 1000
 C21 323 315  16  11  19     316 1000
 C30 900   4  42   7   6      41 1000
 C54 831   4 126  18   0      20 1000
 C22  24  31 508 410   3      23 1000
 C42 514   2 129 306  16      33 1000
 C34 862  65  17  33   8      16 1000
 C51 766  43  69   3 114       6 1000
  C7 747 122   2  55  36      37 1000
 C43 500 104  36 248  18      94 1000
 C49 884  44  14   4  24      30 1000
  C9 147 738  30  27   9      50 1000
  C1 613 274  64   0  11      38 1000
C15p 526  28   1  41 403       1 1000
 C31 526 308  47   0  87      31 1000
 C53  87 450  73  31   2     357 1000
 C23 724 142  18  20  11      86 1000
 C15 207 445  44  25   2     277 1000
 C48  33  11 691  20 220      25 1000
 C33 253  47 385 300   2      12 1000
 TOT 617 139 103  56  43      42 1000
..........
..........

Contribution des variables aux facteurs

       f1   f2   f3   f4   f5  f>5  TOT VARI
  LP  115  243    6  203   26  163  125  245
Lcal  161   42    9  117   56  237  125   49
lsep   52  318  398   63   44   50  125    9
 LTC  153  111   15   48    7  258  125  546
 PGl  146    7    5    1  770   21  125  108
 LTP  150    5   76  313   75   81  125   33
 LTE  169   38    8  118   22  160  125    8
  LF   54  237  483  137    0   31  125    2
     1000 1000 1000 1000 1000 1000 1000 1000
```


Sortir sur papier les résultats de contri pour les analyser en même temps que les biplots.

```
> biplot(forlong.acp) #biplot isométrique ligne sur plan 1-2

1: Biplot isometrique ligne ?
2: Biplot isometrique colonne ?
3: Biplot a_l'hollandaise ?
Selection: 1
```




```
> biplot(forlong.acp,n1=1,n2=3) #biplot isométrique ligne sur
                                   plan 1-3

1: Biplot isometrique ligne ?
2: Biplot isometrique colonne ?
3: Biplot a_l'hollandaise ?
Selection:  1
```



Analyse des variables (résultats contri + biplot):

- représentativité des variables sur plan 1-2 bonne excepté pour lsep et LF bien représentées sur axe 3.
- facteur 1 = facteur taille : à droite les grosses fleurs; à gauche les petites.
- facteurs 2 et 3 sont des facteurs forme.
- comparaison/brévistylés

Sur le plan 1-2, ce qui frappe le plus c'est la position relative de LF et lsep : presque orthogonaux chez les brévi et colinéaires chez longi sur ce plan.
Attention, l'axe 2 représente mal ces 2 variables; il ne faut donc pas conclure trop vite. En effet sur le plan 1-3, leurs positions relatives sont équivalentes à celles observées chez les brévistylés.
On peut remarquer également des corrélations entre variables modifiées entre ces 2 groupes :
chez les brévistylés, on pouvait distinguer 3 goupes de variables : (LTE), (Lcal,LP) et (LTP,PGl,LTC).
chez les longistylés, 2 groupes apparaissent : (LTE,PGl,LTP) et (Lcal,LTC,LP).
Seule la liaison LTP-PGl est maintenue entre les 2 groupes.

Analyse des individus

- représentativité sur les axes : la plupart des individus sont bien représentés sur le plan 1-2. C22, C48 et C33 le sont mal.
- le long de l'axe 1, on distingue les individus à petites fleurs (particulièrement C26 et C54) de ceux à plus grosses fleurs (C51 et C34).
- l'individu C9 se distingue bien sur l'axe 2 (forte lsep, faible LP)

Ajout d'une variable supplémentaire : la longueur des stomates des variétés

Jeu de données = brévistylées

```
> stomat_c(22.07,29.58,25.42,24.41,23.41,24.98,25.26,26.39,23.87,22.25,
           41.52,24.54,25.05,29.32,33.06,23.67,41.16,25.85,27.52,24.39,
           25.87,34.76,26.18,28.61,23.60,26.19,27.40,29.14,22.76,34.82,
           34.69,29.32,31.38,35.87)

> stomat # vecteur longueur des stomates des différentes variétés
           brévistylées

 [1] 22.07 29.58 25.42 24.41 23.41 24.98 25.26 26.39 23.87 22.25 41.52 24.54
[13] 25.05 29.32 33.06 23.67 41.16 25.85 27.52 24.39 25.87 34.76 26.18 28.61
[25] 23.60 26.19 27.40 29.14 22.76 34.82 34.69 29.32 31.38 35.87



> forbrev2.acp_varsup(forbrev.acp,stomat)
> round(forbrev2.acp$vsup,3)
       f1     f2     f3     f4     f5    f6     f7   f8
vs1 0.172 -0.346 -0.363 -0.018 -0.413 0.198 -1.037 0.92
attr(, "moy"):
[1] 28.06794
attr(, "sigma"):
[1] 5.029709


> corsup(forbrev2.acp)
Correlation variables sup./facteurs et cumuls

      f1  f2  f3  f4  f5 f>6
CorS 139 184  83   0  68 142
CorM 139 323 406 406 474 617
```


On travaille sur le groupe des brévistylés car c'est dans ce groupe qu'on trouve le plus de variabilité au niveau de la variable supplémentaire étudiée.

Variable supplémentaire : variable de même type que les variables initiales, ajoutée après analyse.

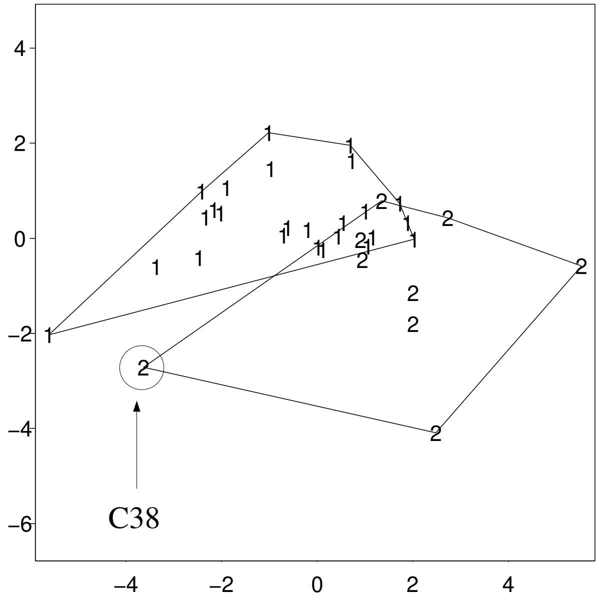
Variable illustrative : variable rajoutant des informations sur les résultats d'analyse.

Ces deux termes (supplémentaire et illustratif) s'opposent au terme de variable active.

Ajout d'une variable illustrative : niveau de ploïdie des variétés

Jeu de données = brévistylées

```
> chrom_c(2,2,2,2,2,2,4,2,2,2,4,2,2,2,4,2,4,2,2,2,2,4,2,2,2,2,2,2,2,
         4,4,2,4,4)
  chrom #vecteur représentant le niveau de ploïdie des
        différentes variétés brévistylées
 [1] 2 2 2 2 2 2 4 2 2 2 4 2 2 2 4 2 4 2 2 2 2 4 2 2 2 2 2 2 2 4 4 2 4 4

> chrom2_chrom/2 #pour que les chiffres correspondent exactement au nb
                de classes on divise par 2 le vecteur précédent
> chrom2
 [1] 1 1 1 1 1 1 2 1 1 1 2 1 1 1 2 1 2 1 1 1 1 2 1 1 1 1 1 1 1 2 2 1 2 2

> plotcl(forbrev.acp$cmpr[,1:2],chrom2) #on place sur le plan
  1-2 du biplot les valeurs de chaque individu pour la
  variable illustrative
```

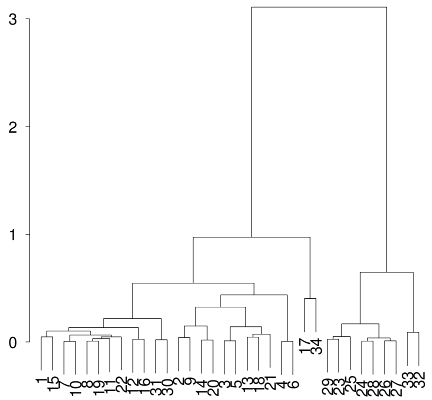


On a une distinction assez nette entre les $2n$ et les $4n$.

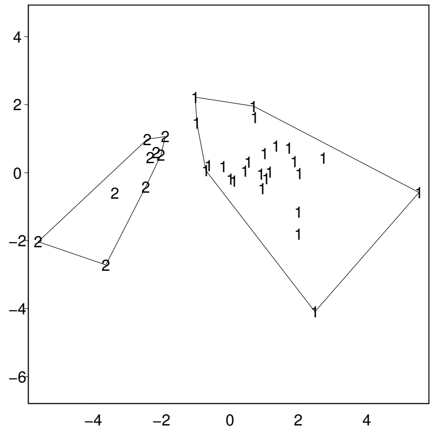
Classification automatique suite à une ACP

Sur le groupe des brévistylés

```
> ddbrev_dist(forbrev.acp$cmpr)
> cahdbrev_cahd(ddbrev)
> plclust(cahdbrev)
```



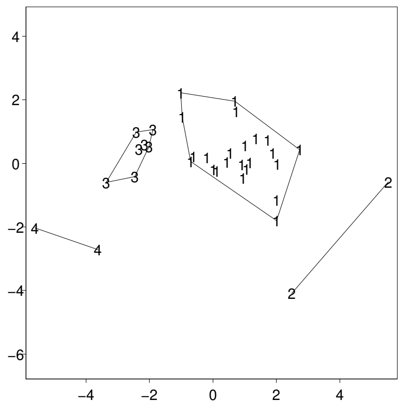
```
> pbrev_cutree(cahdbrev,2)
> plotcl(forbrev.acp$cmpr[,1:2],pbrev)
```



Le but de cette partie n'est pas de tester de façon exhaustive les différentes méthodes de classification automatique, mais de voir ce qu'apporte la classification sur les composantes principales de l'ACP. Plutôt que de partir des données de départ, cette technique permet de travailler avec une métrique adaptée... (voir arguments de A. Carlier et Croquette dans «Nouvelles fonctions pour l'analyse des données multidimensionnelles sous Splus» p30-31).
Nous choisirons le saut de ward comme méthode de classification.

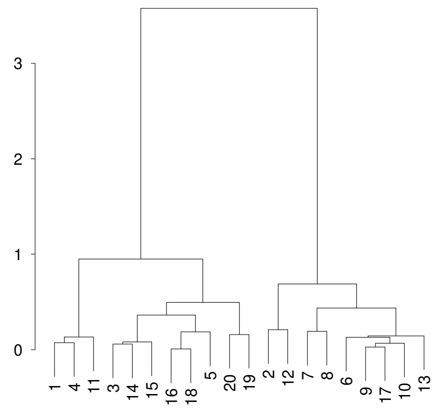
On peut également faire tester aux stagiaires des nombres de clusters différents :
Exemple : 4

```
> pbrev_cutree(cahdbrev,4)
> plotcl(forbrev.acp$cmpr[,1:2],pbrev)
```

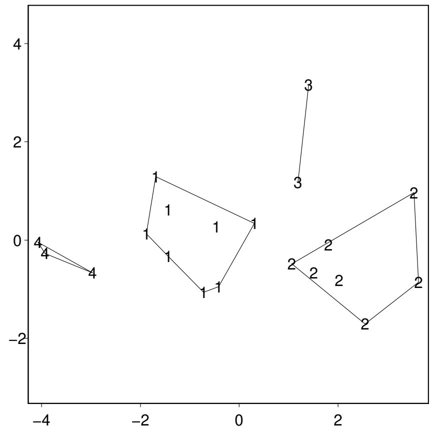


Sur le groupe des longistylés

```
> ddlong_dist(forlong.acp$cmpr)
> cahdlong_cahd(ddlong)
> plclust(cahdlong)
```

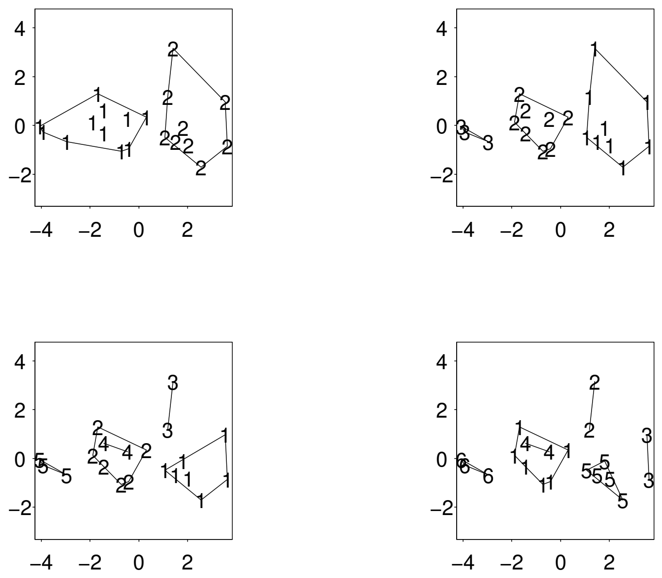


```
> plong_cutree(cahdlong,4) #coupure en 4 groupes
> plotcl(forlong.acp$cmpr[,1:2], plong) #représentation sur le
                                         plan 1-2
```



On peut également faire représenter aux stagiaires d'autres découpages et en discuter

```
> par(mfrow=c(2,2))
> plong2_cutree(cahdlong,2)
> plong3_cutree(cahdlong,3)
> plong5_cutree(cahdlong,5)
> plong6_cutree(cahdlong,6)
> plotcl(forlong.acp$cmpr[,1:2], plong2)
> plotcl(forlong.acp$cmpr[,1:2], plong3)
> plotcl(forlong.acp$cmpr[,1:2], plong5)
> plotcl(forlong.acp$cmpr[,1:2], plong6)
```



Les graphes de la première ligne montrent qu'un seul axe (1) suffit pour différencier les groupes alors que sur les graphes de la deuxième ligne trois dimensions sont nécessaires.

Test sur les résultats de l'ACP du groupe brévistylés de la classification nuées dynamiques ; fonction rkmeans

But : voir les différences de représentations en fonction des itérations
On remarque qu'il y a des groupes assez stables et d'autres beaucoup plus variables

```
> par(mfrow=c(2,2))
> resbrev_rkmeans(forbrev.acp$cmpr,5)

numeros des noyaux
 24 30 32 16 1

> plotcl(forbrev.acp$cmpr[,1:2], resbrev$cluster)
> resbrev_rkmeans(forbrev.acp$cmpr,5)

numeros des noyaux
 8 10 11 28 33

> plotcl(forbrev.acp$cmpr[,1:2], resbrev$cluster)
> resbrev_rkmeans(forbrev.acp$cmpr,5)

numeros des noyaux
 30 6 29 8 16

> plotcl(forbrev.acp$cmpr[,1:2], resbrev$cluster)
> resbrev_rkmeans(forbrev.acp$cmpr,5)

numeros des noyaux
 2 9 1 34 28

> plotcl(forbrev.acp$cmpr[,1:2], resbrev$cluster)
```