# TP3 : Analyse Factorielle Discriminante

*2015-2016*

On considère un jeu de données décrivant trois groupes d'insectes (notés $A$,$B$ et $C$) sur lesquels six mesures anatomiques quantitatives ont été réalisées (notées $X_1, \ldots, X_6$). On veut savoir si ces 6 mesures permettent de retrouver les groupes d'insectes. Pour cela on effectue avec le logiciel R une analyse factorielle discriminante (AFD) ainsi que des calculs de statistiques descriptives (moyenne, variance, corrélation linéaire).

— Les données se trouvent dans le fichier "insectes.rda".

— L'AFD a été implémentée dans R dans deux fonctions `AFD` et `plotAFD` qui se trouvent dans le fichier "AFD_procedures.R".

1. Combien d'axes discriminants peut-on construire avec les données dont on dispose ?
2. Préciser le nombre total d'insectes pris en compte dans l'étude, ainsi que le nombre d'insectes dans chaque groupe.
3. Quel est le centre de gravité $g$ de l'ensemble des données ? Donner les centres de gravité $g_A$, $g_B$ et $g_C$ de chacun des 3 groupes.
4. En considérant l'ensemble des données, quelle est la variable la plus dispersée ? la moins dispersée ?
5. Faire une ACP des données et visualiser les insectes et les variables sur le premier plan factoriel. Commentez. L'ACP vous semble-t-elle une bonne méthode pour discriminer les trois groupes d'insectes ?
6. Faire maintenant un AFD. Au vu des pouvoirs discriminants, combien d'axes semblent nécessaires pour avoir une bonne discrimation des groupes ?
7. Visualisez les insectes sur le premier plan de l'AFD et les variables sur le cercle des corrélations correspondant. Commentez.
8. A votre avis, les données ont-elles été centrées avant de faire l'AFD ?
9. Calculer les coordonnées de la projection des centres de gravité $g_A$, $g_B$ et $g_C$ (des données centrées) sur le premier axe discriminant. Vérifiez qu'il s'agit des moyennes de la première variable discriminante dans chaque groupe.
10. On considère que la première variable disciminante est un bon score. Calculez le score d'un nouvel insecte pour lesquel $X = (193, 131, 55, 160, 16, 102)$ sur cette première variable discriminante.
11. Proposez des seuils pour construire une régle de décision. A votre avis, quel sera le taux d'erreur avec cette règle de décision sur les données ? A quelle classe appartient le nouvel insecte de la question précédente avec cette règle de décision ?

12. Construire le score et la règle de décision de l'AFD sur un échantillon aléatoire de 50 insectes (formant un échantillon d'apprentissage). Calculer les scores des 24 autres insectes (l'échantillon test) et classer ces 24 insectes avec la régle de décision construite sur l'ensemble d'apprentissage. Quel est le taux d'erreur ?

13. On veut maintenant comparer l'approche "anglo-saxonne" et l'approche "francophone" de l'AFD. Pour cela, utilisez l'argument type="FR" et type="GB" dans la fonction AFD du code R. Comparez numériquement les valeurs propres des deux approches. Comparez numériquement puis visuellement les variables discriminantes.

14. On peut utiliser les fonctions **lda** et **predict. lda** du package **MASS** pour d'effectuer une analyse canonique discriminante (terminologie anglo-saxonne pour une AFD).

    (a) Pour obtenir les facteurs et les variables discriminantes on tape le code :

    ```
    require(MASS)
    res <- lda(type~.,insectes)
    pred <- predict(res)
    res$scaling
    pred$x
    ```

    Les arguments "scaling" et "x" des fonctions **lda** et **predict. lda** fournissent respectivement les facteurs et les variables discriminantes.

    (b) Vérifier dans R que l'on retrouve bien les mêmes résultats avec :

    ```
    res2 <- AFD(X,y,type="GB")
    res2$U
    res2$S
    ```

### Résultats des statistiques descriptives :

```
Taille de l'ensemble des groupes :
----------------------------------
[1] 74

Moyennes pour l'ensemble des groupes :
--------------------------------------
    [,1]
X1 177.3
X2 124.0
X3  50.4
X4 134.8
X5  13.0
X6  95.4

Matrice de variances pour l'ensemble des groupes :
------------------------------------------
        X1    X2    X3      X4     X5      X6
X1  853.41  6.48 -7.64 -100.60  48.56 -237.29
X2    6.48 70.96 15.50   48.63  -2.19   58.43
X3   -7.64 15.50  7.47   16.66  -1.82   20.04
X4 -100.60 48.63 16.66  105.69  -5.49  114.60
X5   48.56 -2.19 -1.82   -5.49   4.53  -14.47
```

```
X6 -237.29 58.43 20.04  114.60 -14.47  201.86

Matrice de correlations pour l'ensemble des groupes :
-----------------------------------------------------
      X1    X2    X3    X4    X5    X6
X1  1.00  0.03 -0.10 -0.33  0.78 -0.57
X2  0.03  1.00  0.67  0.56 -0.12  0.49
X3 -0.10  0.67  1.00  0.59 -0.31  0.52
X4 -0.33  0.56  0.59  1.00 -0.25  0.78
X5  0.78 -0.12 -0.31 -0.25  1.00 -0.48
X6 -0.57  0.49  0.52  0.78 -0.48  1.00


Taille de l'echantillon du groupe ' A ' :
----------------------------------------
[1] 21


Moyennes pour le groupe ' A ' :
------------------------------

X1 183.1
X2 129.6
X3  51.2
X4 146.2
X5  14.1
X6 104.9

Matrice de variances pour le groupe ' A ' :
-------------------------------------------
       X1    X2    X3    X4    X5    X6
X1 140.47 63.46 17.64 14.36 -4.96 13.54
X2  63.46 48.81 11.00  2.36 -1.73  2.95
X3  17.64 11.00  4.75  5.57 -0.50  5.22
X4  14.36  2.36  5.57 30.15 -0.92 14.88
X5  -4.96 -1.73 -0.50 -0.92  0.75 -1.89
X6  13.54  2.95  5.22 14.88 -1.89 36.41


Taille de l'echantillon du groupe ' B ' :
----------------------------------------
[1] 22

Moyennes pour le groupe ' B ' :
------------------------------

X1 138.2
X2 125.1
X3  51.6
X4 138.3
X5  10.1
X6 106.6

Matrice de variances pour le groupe ' B ' :
-------------------------------------------
      X1    X2    X3    X4    X5    X6
X1 83.36 42.52 19.59 18.30 -0.70 14.59
X2 42.52 69.72 14.99 13.38 -0.37 20.26
X3 19.59 14.99  7.70  7.84 -0.28  4.74
X4 18.30 13.38  7.84 16.38 -0.48  7.57
X5 -0.70 -0.37 -0.28 -0.48  0.90  0.26
X6 14.59 20.26  4.74  7.57  0.26 32.70
```

```
Taille de l'echantillon du groupe ' C ' :
----------------------------------------
[1] 31

Moyennes pour le groupe ' C ' :
------------------------------

X1 201.0
X2 119.3
X3  48.9
X4 124.6
X5  14.3
X6  81.0

Matrice de variances pour le groupe ' C ' :
-------------------------------------------
       X1    X2    X3    X4    X5    X6
X1 214.97 61.35 21.87 29.39  4.23 28.52
X2  61.35 42.73  7.65 11.44  0.33 11.10
X3  21.87  7.65  5.34  5.50  0.01  4.10
X4  29.39 11.44  5.50 20.68 -0.32 11.32
X5   4.23  0.33  0.01 -0.32  1.17  1.23
X6  28.52 11.10  4.10 11.32  1.23 77.16
```

## Resultats numériques de l'AFD :

```
Liste des pouvoirs discriminants :
----------------------------------
[1] 0.947 0.795

Matrice des facteurs discriminants :
------------------------------------
      u1    u2
X1 -0.02  0.01
X2  0.01  0.02
X3  0.03 -0.13
X4  0.02  0.09
X5 -0.07  0.24
X6  0.01  0.01

Matrice des variables discriminantes :
--------------------------------------
      s1    s2
1   0.08  1.82
2   0.25  1.48
3  -0.12  1.31
4   0.08  1.67
5   0.41  1.78
6   0.17  1.22
7   0.23  1.54
8   0.12  1.15
9   0.52  1.11
10  0.05  1.34
11  0.17  0.82
12  0.36  2.12
........
........
```

```
67 -1.21 -0.91
68 -1.18 -1.43
69 -0.47 -0.45
70 -0.83 -0.42
71 -1.27 -0.38
72 -0.79 -0.51
73 -0.86 -1.01
74 -0.50 -0.09

Matrice descorrelations avec les variables discriminantes :
--------------------------------------
      s1   s2
X1 -0.90 0.26
X2  0.34 0.43
X3  0.45 0.17
X4  0.65 0.70
X5 -0.80 0.48
X6  0.82 0.37
```

## Graphiques de l'AFD :