SY09 - TP02
Analyse factorielle d'un tableau de distances, classification automatique

Nicolas Szewe - Marie Chatelin

April 23, 2015

1 Introduction

Ce rapport a comme objectif la mise en pratique des différentes méthodes de classification automatique. La méthode d'Analyse Factorielle d'un Tableau de Distances, la méthodes des centres mobiles et les classifications hierarchiques seront ainsi étudiées.

2 Analyse factorielle d'un tableau de distances

2.1 Partie Théorique

En utilisant la matrice définie à partir de X ou calculée à partir de D2, nous obtenons la matrice W suivante :

W =

$$\begin{pmatrix} 13.50 & -8.78 & -16.56 & 16.25 & 11.07 & -13.81 & 12.44 & -14.12 \\ -8.78 & 6.19 & 11.16 & -11.03 & -7.71 & 8.91 & -8.84 & 10.10 \\ -16.56 & 11.16 & 20.63 & -20.31 & -14.00 & 16.88 & -15.87 & 18.07 \\ 16.25 & -11.03 & -20.31 & 20.00 & 13.82 & -16.56 & 15.69 & -17.87 \\ 11.07 & -7.71 & -14.00 & 13.82 & 9.63 & -11.25 & 11.00 & -12.56 \\ -13.81 & 8.91 & 16.88 & -16.56 & -11.25 & 14.13 & -12.62 & 14.32 \\ 12.44 & -8.84 & -15.87 & 15.69 & 11.00 & -12.62 & 12.63 & -14.43 \\ -14.12 & 10.10 & 18.07 & -17.87 & -12.56 & 14.32 & -14.43 & 16.50 \end{pmatrix}$$

Vérifions maintant que W est semi-définie positive en calculant les valeurs propres de W.
Les valeurs propres de W sont :

lambda =

$$\begin{pmatrix} 111.46 & 1.76 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Nous avons ramener à 0 les valeurs négatives à 10e-15 car ceci n'était pas significatif dans notre cas. La matrice W est donc bien semi-définie postive car toutes ses valeurs propres sont positives ou nulles.

La matrice $W_(1/8)$ est :

$W_(1/8)$ =

$$\begin{pmatrix} 1.688 & -1.097 & -2.07 & 2.032 & 1.383 & -1.726 & 1.555 & -1.765 \\ -1.097 & 0.774 & 1.395 & -1.378 & -0.964 & 1.114 & -1.105 & 1.262 \\ -2.07 & 1.395 & 2.579 & -2.539 & -1.75 & 2.11 & -1.984 & 2.258 \\ 2.032 & -1.378 & -2.539 & 2.5 & 1.727 & -2.07 & 1.961 & -2.234 \\ 1.383 & -0.964 & -1.75 & 1.727 & 1.204 & -1.406 & 1.375 & -1.57 \\ -1.726 & 1.114 & 2.11 & -2.07 & -1.406 & 1.766 & -1.578 & 1.79 \\ 1.555 & -1.105 & -1.984 & 1.961 & 1.375 & -1.578 & 1.579 & -1.804 \\ -1.765 & 1.262 & 2.258 & -2.234 & -1.57 & 1.79 & -1.804 & 2.063 \end{pmatrix}$$

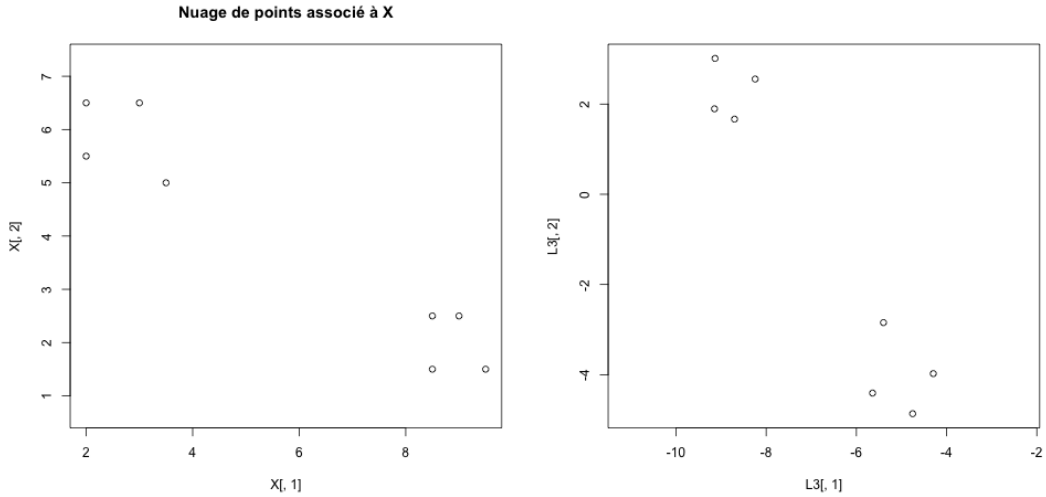


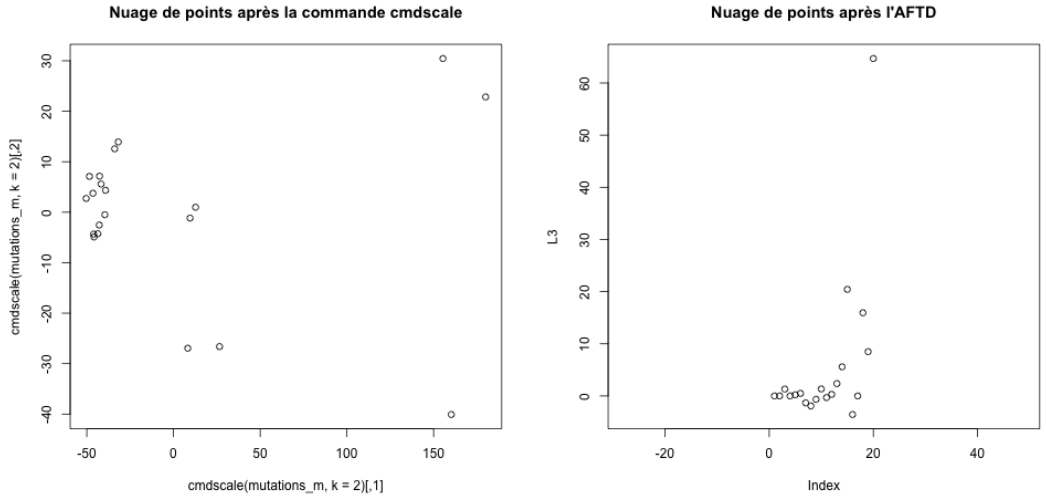
Table 1 : Comparaison nuage du tableau initial et de celui associé à la représentation fournie par l'AFTD

On remarque que ces representations dans le premier plan factoriel sont ne sont pas sur la même échelle. Cependant la dispersion des points dans l'espace est la même, à la symétrie près. Globalement l'AFTD pour ces données a une bonne représentation. On constate cependant, qu'il y a des valeurs propres négatives assez grandes.

```
AFTD = function(D, k = dim(D)[1]){
n = dim(D)[1]
In = diag(1, n)
Un = matrix(data = 1, nrow = n, ncol = n)
Qn = (In - (Un/n))
D2 = D2̂
W =((-1/2)*Qn%*%(D2)%*%Qn)*(1$n)
L1 = sqrt(n) * round((eigen(W/n)$vectors))[,1:k]
Nous ne prenons que les valeurs propres positives, celles négatives sont source
d'erreur L = abs(eigen(W/n)$values)
L3 = L1%*%sqrt(L)
plot(L3, asp = 1, main = "Nuage de points après l'AFTD")
list(L3, Quality = 100*(L[1] + L[k])/sum(L))
}
```

2.2 Analyse des données mutations

Nous allons appliquer l'AFTD à ce tableau de dissimilarités afin d'obtenir l'expression de chacune des espèces dans un espace de faible dimension.



Le diagramme qui permet de regarder la proximité entre les dissimilarités et les distances est le diagramme de Shepard. Si on a des bons résultats, les points sont observés sur la diagonale.
Au départ on travaille avec un tableau de dissimilarité. On a obtenu par des méthodes d'analyse, on ne sait pas si les valeurs sont vraiment des distances mais on considère que ce sont des dissimilarités.
Plus on prend de dimension, plus on a des chances de trouver une configuration et que dans ce nouvel espace, les

distances soient proches des dissimilarités. Sur le diagrame, nous mettons les distances en Y et les dissimilarités en X. Nous voulons alors obtenir des points le long de la diagonale afin de pouvoir conclure sur la fiabilité du diagramme de Shepard. C'est seulement à ce moment là que l'on peut conclure sur les résultats dans le nouvel espace. Malheureusement, nous n'avons pas réussi à générer un diagramme correcte.

3 Méthode des centres mobiles

3.1 Données Iris

3.1.1 Première approche

Nous représentons nos partitions sur le premier plan factoriel.

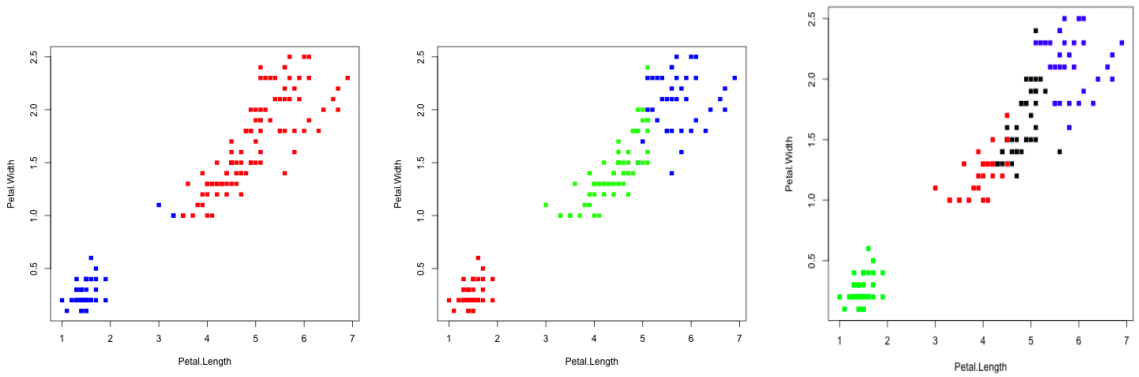


Table 2 : Representation de nos partitions sur le premier plan factoriel K = 2,3,4

Nous remarquons que la classification en deux partitions nous donne deux groupes bien distinct. On remarque cependant quelques points proche du centre du cluster auquel ils n'appartiennent pas. Notamment deux points bleus dans le cluster rouge. On note que la séparation en trois clusters "répare" cette erreur.

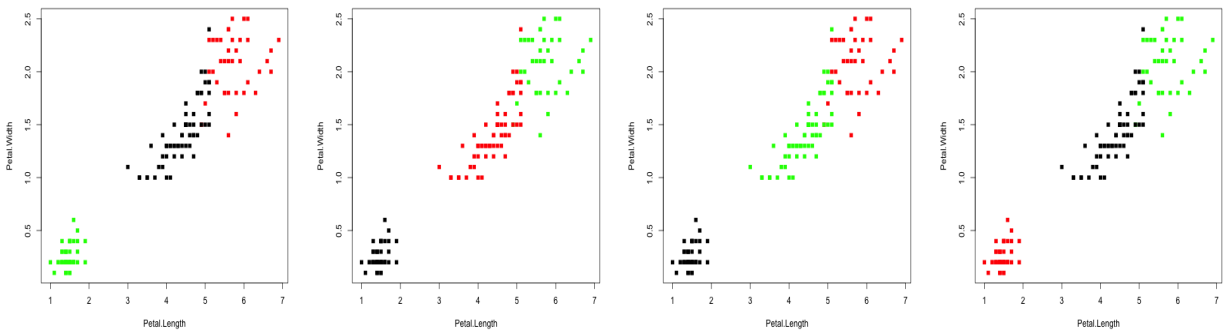
Pour la partition en 4 groupes, nous remarquons que les groupe rouge et noir se mélangent. Certains points sont donc plus proche du centre mobiles d'une classe auquel ils n'appartiennent pas.

Ces erreurs de classifications peuvent provenir de la manière dont les classifications par les centres mobiles sont effectuées. En effet, le carre de la distance Euclidienne étant utilisé comme critère de proximité, nous supposons que les variables ont approximativement la même variance. Or les variances des données Iris ne sont sensiblement pas les mêmes.

3.1.2 Stabilité de l'algorithme

Nous cherchons maintenant à étudier la stabilité du résultat de la partition.

Nous remarquons que si on refait plusieurs fois le calcul des classes, nous n'obtenons jamais le même résultat. En effet, les premiers centres sont choisis au hasard. En faisant tourner l'algorithme plusieurs fois, nous tendons cependant vers des résultats similaires.



En changeant les centres, nous avons un problème d'identifiabilité (en apprentissage non supervisé, on ne connait pas les vraies classes). Il faut donc prouver que les deux partitions sont les mêmes (au nommage près). On remarque que les classes sont à peu de chose près les mêmes mais que leur couleur donc leur ID ne sont pas toujours les mêmes.

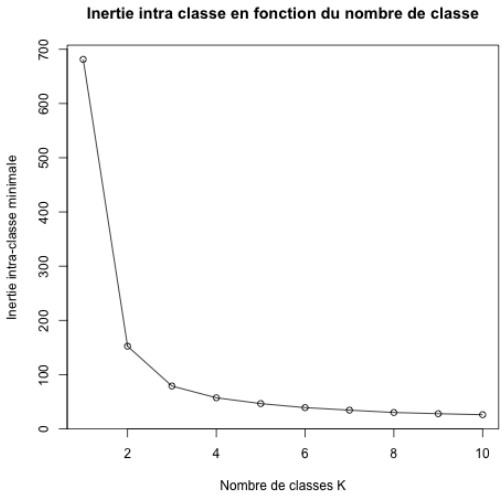
3.1.3 Recherche du K optimal

Nous allons maintenant chercher le K optimal de classe pour notre jeu de donnée. Pour ce faire, nous effectuons 100 fois la méthode pour chaque K entre 2 et 9. Nous allons ainsi chercher, à minimiser l'inertie intra-classe.

Les valeurs d'inerties intra-classe sont ainsi les suivantes :

VI =

$$\begin{pmatrix} NbClasse & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ Inertie & 681.4 & 152.3 & 78.9 & 57.2 & 46.4 & 39.1 & 34.4 & 30.1 & 27.8 & 26.0 \end{pmatrix}$$



En utilisant la méthode du coude, le nombre de classe optimal choisi est donc de 3. Cette valeur, nous semble cohérente étant le nombre de variétés d'iris présentes dans nos données.

3.1.4 Comparaison partition obtenue par les centres mobiles avec la partition réelle des iris en trois groupes

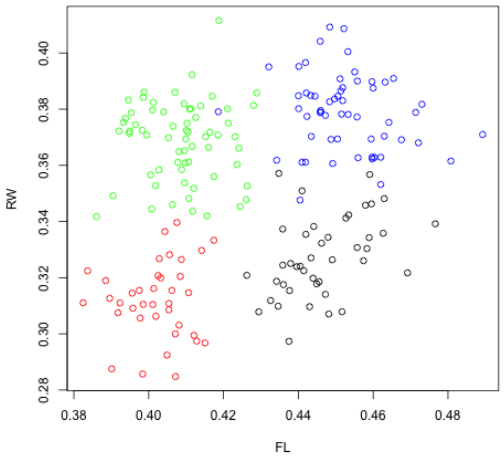
Pour comparer ces partitions, nous pouvons utiliser un tableau de contingence.

$$\begin{pmatrix} & 1 & 2 & 3 \\ Setosa & 0 & 0 & 50 \\ Versicolor & 3 & 47 & 0 \\ Virginica & 38 & 12 & 0 \end{pmatrix}$$

Nous remarquons que les partitions obtenues par les centres mobiles ne sont pas strictement les mêmes. Cette différence est surtout présente au niveau du K = 2 pour laquelle les espèces Versicolor et Virginica se mélangent.

3.2 Données Crabs

En utilisant l'algorithme des centres mobiles sur les données Crabs, pour K = 4, nous obtenons le graphe suivant :



Nous avons choisi de représenter notre classification sur le plan (FL,RW) car c'est sur ce plan que la représentation était la plus significative visuellement.

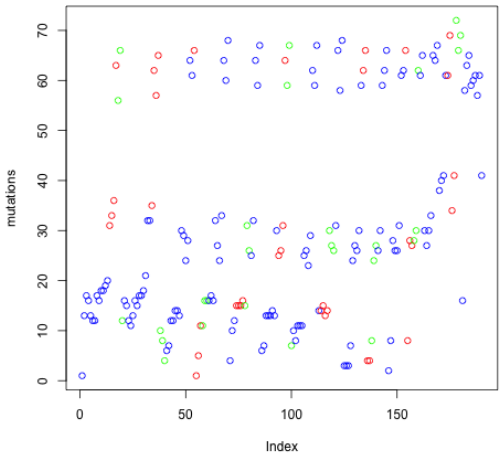
En réalisant un tableau de contingence, nous pouvons comparer nos partitions à celle des crabes suivant leur espèce ou leur sexe. Nous obtenons ainsi le tableau suivant :

$$\begin{pmatrix} & 1 & 2 & 3 & 4 \\ BF & 0 & 0 & 0 & 50 \\ OF & 49 & 0 & 0 & 1 \\ BM & 0 & 0 & 37 & 13 \\ OM & 8 & 42 & 0 & 0 \end{pmatrix}$$

On remarque que la classification répartie plutôt bien les espèces féminines mais est moins précise pour les espèces masculines.

3.3 Données Mutation

Nous avons representer nos données dans le premier plan factoriel après avoir réalisé une partition en 3 classes.



En réalisant l'expérience 10 fois, on remarque que les résultats restent relativement similaires. Dans le cadre de l'étude mutation, nous avons suivi un exemple pour lequel la partition n'est pas satisfaisante puisque les partitions ne sont pas du tout distinctes.