SORBONNE UNIVERSITES

21 avril 2017

TP 2 - Classification automatique

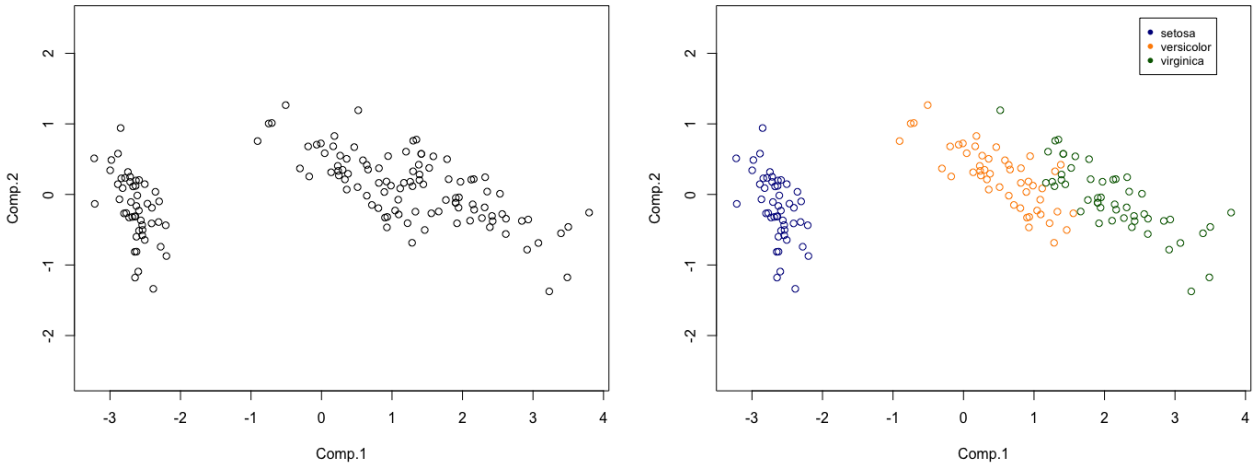
UV : **SY09**
Branche : **Génie Informatique**
Filière : **Fouille de Données et Décisionnel**
Auteurs : **LU Han - HAMONNAIS Raphaël**

Table des matières

1. Visualisation des données

1.1 Visualisation des données Iris

FIGURE 1.1 – Représentation des données `Iris` dans le premier plan factoriel après ACP



(a) Données `Iris` dans le premier plan factoriel sans tenir compte de l'espèce

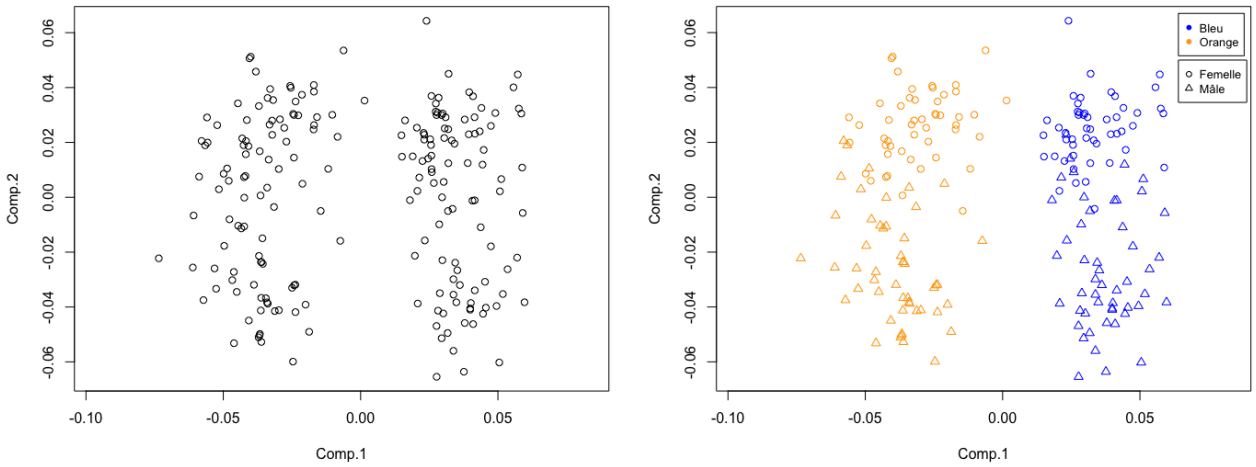
(b) Données `Iris` dans le premier plan factoriel en tenant compte de l'espèce

» Affichage dans le premier plan factoriel sans tenir compte de l'espèce
  — On observe deux groupes bien distincts (voir Figure 1.1a).

» Affichage dans le premier plan factoriel en tenant compte de l'espèce
  — On voit qu'un des deux groupes précédemment observé est en fait constitué de deux espèces différentes (voir Figure 1.1b) ;
  — On obtient donc deux informations précieuses :
    — Les méthodes de classification géométriques tendront à nous donner deux classes ;
    — On sait que les données contiennent en réalité trois classes bien distinctes quand le facteur discriminant est l'espèce ;
    — Il faudra donc faire attention à ce qu'on cherche à obtenir : une nouvelle classification en X classes sans tenir compte de l'espèce, et auquel cas on obtiendra surement deux classes. Ou bien une classification en fonction de l'espèce et alors il faudra spécifier qu'on cherche à obtenir trois classes.

1.2 Visualisation des données Crabs

FIGURE 1.2 – Représentation des données `Crabs` dans le premier plan factoriel après ACP



(a) Données `Crabs` dans le premier plan factoriel sans tenir compte de l'espèce ou du sexe

(b) Données `Crabs` dans le premier plan factoriel en tenant compte de l'espèce et du sexe

» Affichage dans le premier plan factoriel sans tenir compte de l'espèce ou du sexe
  — On observe deux groupes bien distincts (voir Figure 1.2a).

» Affichage dans le premier plan factoriel en tenant compte de l'espèce et du sexe (voir Figure 1.2b)
  — On constate que les deux groupes observés précédemment correspondent à l'espèce des crabes ;
  — On voit aussi apparaître deux autres groupes au sein des premiers qui délimitent le sexe ;
  — On va donc chercher à faire une classification à 4 classes ;
  — On note quand même que la délimitation entre les sexes est plus floue que celle entre les espèces.

1.3 Visualisation des données Mutation

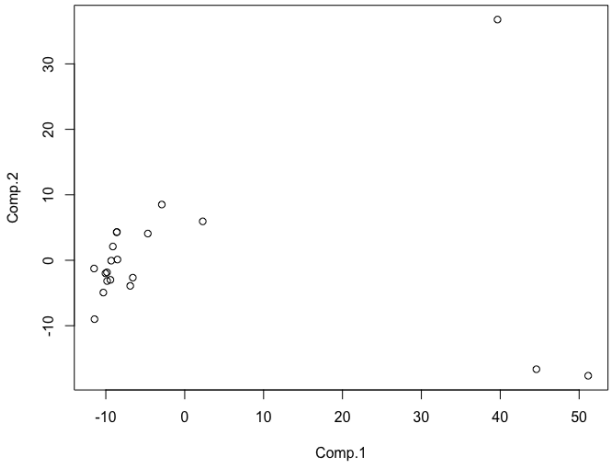
Les données `Mutation` représentent par le biais d'une matrice de dissimilarités les liens entre espèces : plus la distance (dissimilarité) est faible, plus les espèces sont proches.

Nous allons effectuer une Analyse Factorielle de Tableau de Distance (AFTD). On rappelle que l'AFTD peut être vue comme un équivalent de l'ACP pour des données se présentant sous la forme d'un tableau $n \times n$ de dissimilarités δ_{ij} entre n individus $(i,\ j \in \{1, ..., n\})$: elle calcule une représentation multidimensionnelle de ces individus (dont le tableau de dissimilarités ne donne qu'une description implicite) dans un espace euclidien de dimension $p \leqslant n$. Cette représentation est exacte lorsque les dissimilarités sont des distances euclidiennes, ce qui n'est pas toujours le cas.

Après sélection d'un certain nombre de variables, la qualité de la représentation peut être évaluée numériquement par un critère similaire au pourcentage d'inertie de l'ACP, ou graphiquement au moyen d'un diagramme de Shepard : sur ce graphique, la distance $d_{ij} = d(x_i, x_j)$ entre les représentations de x_i et x_j déterminées par l'AFTD est représentée en fonction de la dissimilarité initiale δ_{ij}, pour chaque couple d'individus (x_i, x_j).

1.3.1 Données Mutation dans le premier plan factoriel après AFTD

FIGURE 1.3 – Représentation euclidienne des données `Mutation` en deux dimensions par AFTD



Cette représentation a le mérite de nous permettre d'appréhender plus facilement le tableau de dissimilarités : on voit clairement que beaucoup des espèces sont proches les unes des autres. Certaines très proches. On voit aussi que trois d'entre elles sont particulièrement éloignées.

On pourrait former deux classes : la première regroupant l'ensemble des points proches, et la seconde les trois points éloignés. Ou bien trois, si on décide que l'espèce tout en haut à droite du graphique est trop loin pour être intégrée à une classe. Il est aussi tout à fait possible de subdiviser la première classe d'espèces proches les unes des autres en plusieurs classes plus petites.

1.3.2 Analyse de la qualité de la représentation par AFTD

Certaines des valeurs propres (inertie expliquée de la composante principale correspondante) sont négatives. Calculer le pourcentage d'inertie expliquée demande alors de faire un choix : transformer les valeurs propres négatives en leur inverse positif (valeur absolue) ou bien ne tenir compte que des valeurs propres positives. Nous avons ici effectué les calculs avec les deux possibilité afin de comparer les résultats (voir Tableau 1.1 et Tableau 1.2).

TABLE 1.1 – Inertie avec valeurs absolues

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Pourcentage d'inertie expliquée	52.71	16.01	10.94	6.72	4.78
Pourcentage cumulé d'inertie expliquée	52.71	68.72	79.67	86.38	91.16

TABLE 1.2 – Inertie avec valeurs positives seulement

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Pourcentage d'inertie expliquée	53.43	16.23	11.09	6.81	4.84
Pourcentage cumulé d'inertie expliquée	53.43	69.66	80.75	87.56	92.40

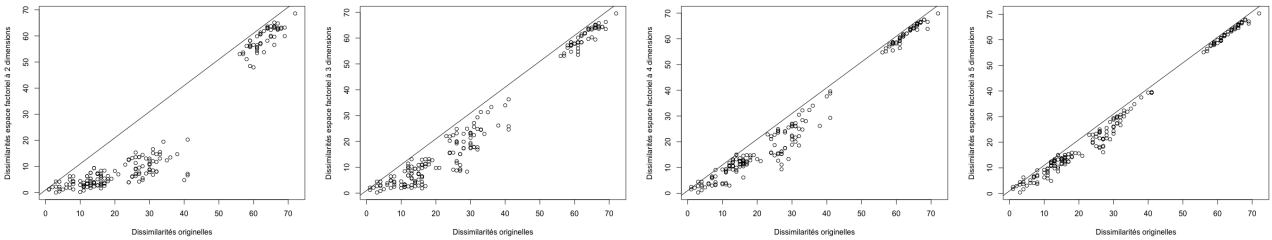
On remarque que les pourcentages d'inertie expliquée ne varient pas énormément entre les deux choix de calcul. Cela s'explique par la faible importance des valeurs propres négatives dans ce jeu de données.

Pour ce qui est de la qualité de la représentation, on voit très vite qu'avec simplement deux dimensions, on n'obtient qu'environ 69% d'inertie expliquée. C'est relativement peu pour une ACP, mais cela peut être

suffisant pour une représentation simple des données. Tenir compte de trois, quatre ou cinq dimensions nous permet à chaque fois de mieux représenter des données, avec respectivement 80%, 87% et 92% d'inertie cumulée.

On remarque aussi cette augmentation de qualité de représentation des données initiales lorsqu'on regarde les diagrammes de Shepard :

FIGURE 1.4 – Diagrammes de Shepard pour les dimensions 2 à 5



(a) Diagramme de Shepard (2 dimensions)

(b) Diagramme de Shepard (3 dimensions)

(c) Diagramme de Shepard (4 dimensions)

(d) Diagramme de Shepard (5 dimensions)

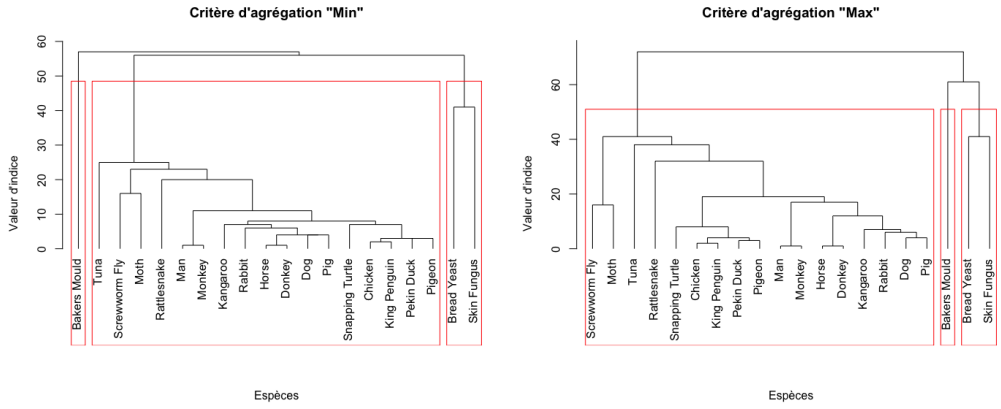
L'axe des abscisses représente les dissimilarités originelles avant AFTD et l'axe des ordonnées les distances euclidiennes entre les observations sur le nouvel espace factoriel obtenu par l'AFTD.

Plus le nombre de dimensions augmente, plus les valeurs des dissimilarités avant et après AFTD s'approchent de la droite $y = x$, c'est à dire une dissimilarité initiale égale à la distance entre les individus représentés dans l'espace euclidien à k dimensions défini par l'AFTD.

2. Classification hiérarchique

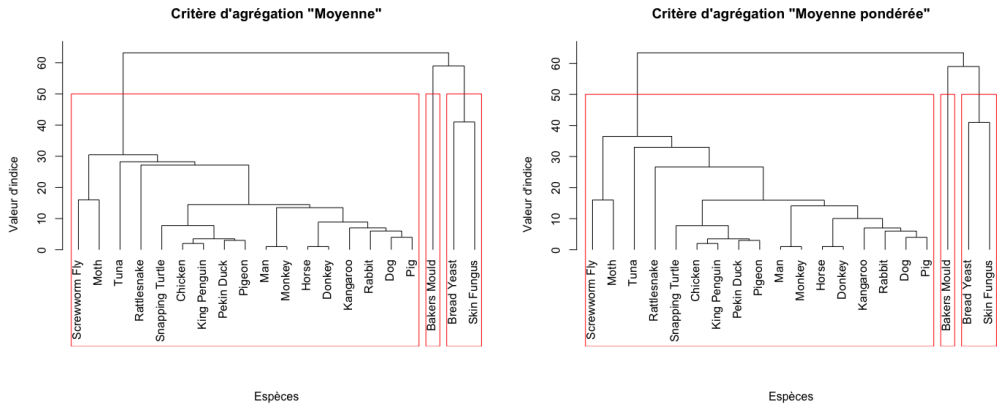
2.1 Classification hiérarchique ascendante sur les données Mutation

Remarque : certains critères d'agrégation sont « pondérés ». Cela signifie que les classes sont considérées comme étant de poids équivalents, quel que soit leur effectif.



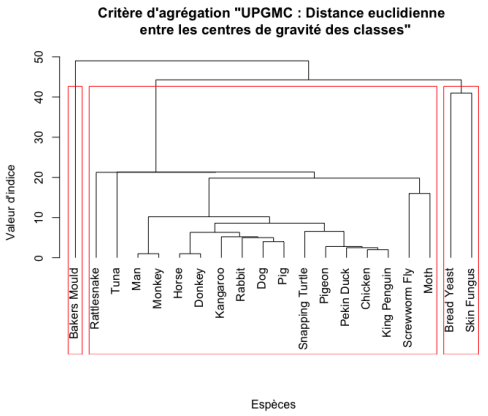
(a) Critère agrégation Min (b) Critère agrégation Max

FIGURE 2.1 – Classification avec les critères d'agrégation Min et Max

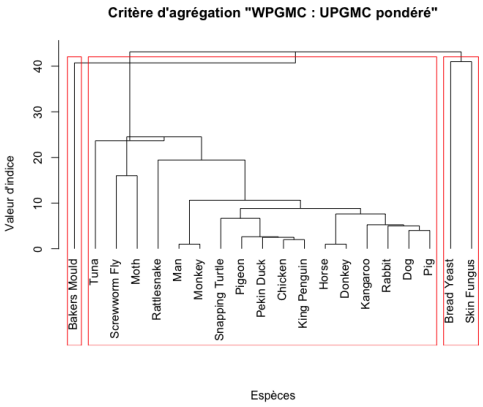


(a) Critère agrégation Moyenne (b) Critère agrégation Moyenne pondérée

FIGURE 2.2 – Classification avec les critères d'agrégation Moyenne et Moyenne pondérée



(a) Critère agrégation distance inter-classe



(b) Critère agrégation distance inter-classe pondérée

FIGURE 2.3 – Classification avec les critères d'agrégation de distance inter-classe (distance entre centres de gravité des classes)

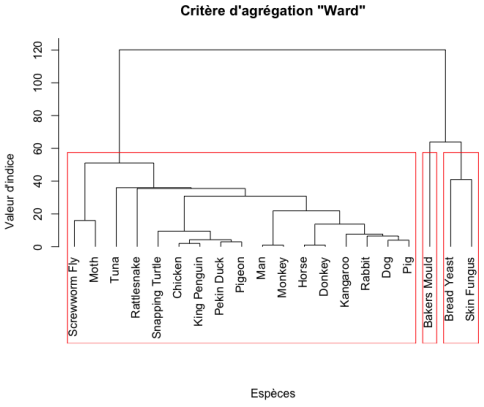


FIGURE 2.4 – Classification avec le critère d'agrégation de Ward

La première chose qu'on remarque est que, lorsqu'on divise en trois classes tel que nous le suggérait la représentation graphique des données après AFTD (cf. sous-section 1.3.1), tous les critères d'agrégation donnent les mêmes classes. A savoir {*Bakers Mould*}, {*Bread Yeast, Skin Fungus*}, et le reste dans la dernière classe.

Il est aussi intéressant de remarque que deux des critères d'agrégation ne sont pas monotones (cf. Figure 2.3a et Figure 2.3b). On voit en effet que l'indice n'est pas décroissant.

2.2 Classification hiérarchique sur les données Iris

2.2.1 Classification hiérarchique ascendante

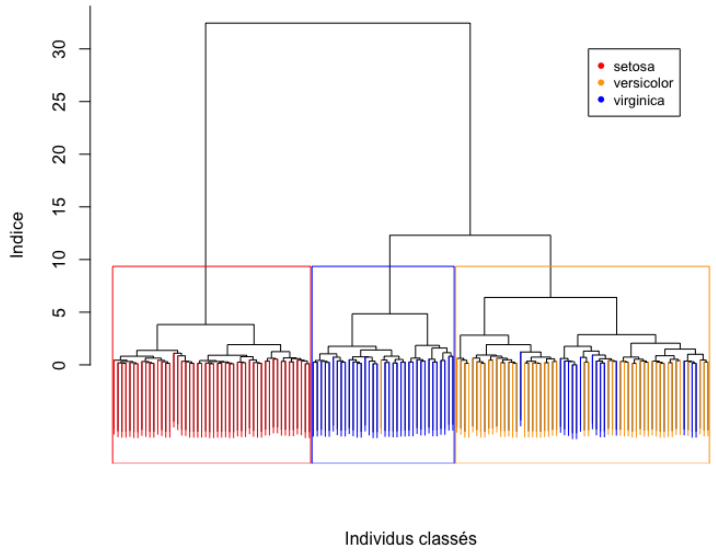


FIGURE 2.5 – Iris : classification hiérarchique ascendante

La classification identifie très bien les individus de l'espèce *setosa*. Par contre, elle tend à confondre certains individus des espèces *versicolor* et *virginica*. Compte tenu de la représentation graphique obtenue après ACP (cf. Figure 1.1b), ce n'est pas étonnant, ces deux espèces étant très proches sur le graphe, contrairement à *setosa* qui est bien distincte.

2.2.2 Classification hiérarchique descendante

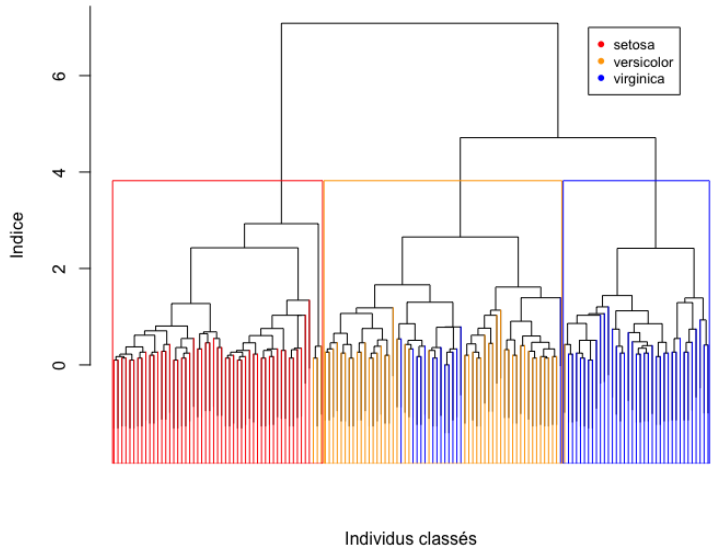


FIGURE 2.6 – Iris : classification hiérarchique descendante

La classification hiérarchique descendante confond elle aussi certains des individus des espèces *versicolor* et *virginica*. Afin de savoir laquelle des deux méthodes reste la plus précise, nous utilisons l'indice de Rand corrigé. Le principe est de sélectionner l'ensemble des paires possibles d'individus au sein de chaque partition et

comparer leur classement au classement initial qui est connu. La valeur de cet indice est comprise entre 0 et 1 : plus elle se rapproche de 1, plus la classification obtenue est proche de la classification initiale.

Valeurs obtenues :

- Classification ascendante : 0.73
- Classification descendante : 0.69

On en déduit que la classification ascendante est meilleure pour ce jeu de données, bien que de peu.

3. Méthode des centres mobiles (K-Means)

3.1 Données Iris

3.1.1 Partition en $K \in 2, 3, 4$

3.1.2 Étude de la stabilité du résultat

3.1.3 Détermination du nombre de classes optimal

3.1.4 Comparaison de la classification obtenue avec la classification initiale

3.2 Données Crabs

3.2.1 Classification en 2 classes

3.2.2 Classification en 4 classes

3.3 Données Mutations

3.3.1 Classification en 3 classes

3.3.2 Étude de la stabilité du résultat