Réduction de dimensionnalité et techniques de clustering

Dr. Matthieu cisel

September 2024

1 Introduction

Dans ce cours centré sur l'apprentissage non supervisé, nous allons porter la focale sur la réduction de dimensionnalité et le clustering. Nous explorerons en particulier différents aspects de la création et de l'interprétation de l'analyse en composantes principales, et de l'analyse des correspondances multiples. En ce qui concerne les techniques de clustering, nous en verrons trois : le k-means, la classification ascendante hiérarchique, et DBSCAN. Sachez que Python est nettement moins riche que R en termes de fonctionnalités. Nous préciserons les étapes que seuls les utilisateurs de R seront invités à réaliser.

Vous serez guidé pas à pas. Le premier projet se fonde sur le jeu de données Iris, classiquement utilisé dans les cours sur l'analyse en composantes principales. Dans un second projet, nous fournissons moins de guidance, vous travaillerez sur des données classiques d'un décathlon, toujours sur une ACP, mais avec une focale plus grande sur le clustering. Dans le troisième projet, la focale porte sur l'analyse des correspondances multiples; nous avons voulu explorer ce faisant une seconde technique de réduction de dimensionnalité. Le jeu de données mobilisé est artificiel. Créé par nos soins, il décrit des profils d'une application de dating. Nous faisons en dernier lieu un ultime exercice de clustering, fondé sur DBSCAN.

La maîtrise des commandes requises pour réaliser les exercices listés ici passe par le suivi d'un cours de Datacamp. Il s'agit pour Python du cours suivant, et pour R du cours suivant. Le cours sur Python déborde un peu par rapport aux objectifs du projet, contrairement au cours sur R. Des recherches sur Internet seront nécessaires sur certaines commandes.

2 Réduction de dimensionnalité

2.1 Projet 1

1. Chargez dans votre environnement le jeu de données IRIS, et représentez en trois dimensions (comme dans la Figure 1) les individus sur les axes suivants : Sepal.Length, Petal.Length, et Sepal.Width

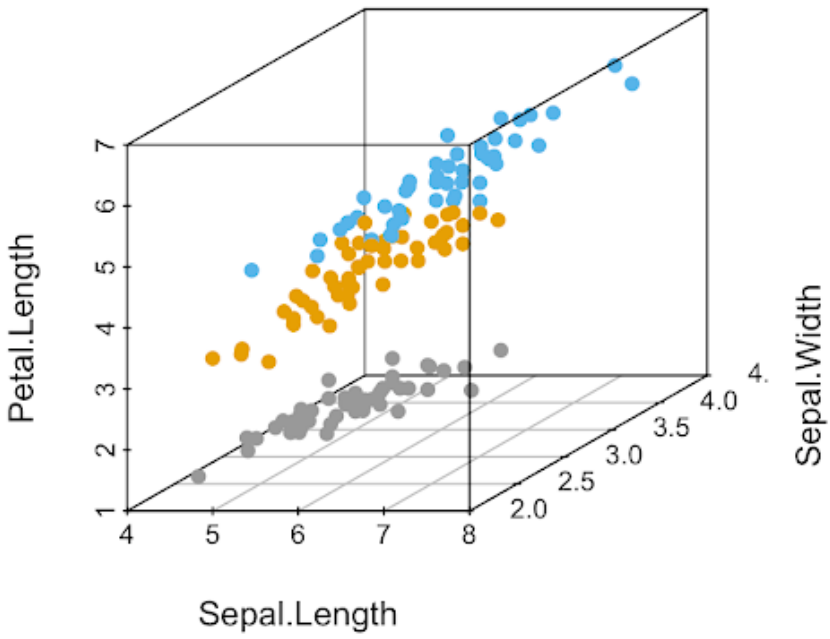

Figure 1: Iris data visualization before dimensionality reduction

2. Répétez l'étape précédente, mais après avoir centré puis réduit les variables correspondantes

3. Réalisez une analyse en composantes principales, puis affichez les individus sur les deux premières composantes. Vous choisirez des couleurs

différentes selon le type d'espèce d'iris considéré, comme en Figure 2. Vous réaliserez cette étape en deux fois : une première fois sans standardisation des données, une seconde fois avec. A partir de maintenant, vous travaillerez toujours avec les variables centrées réduites.

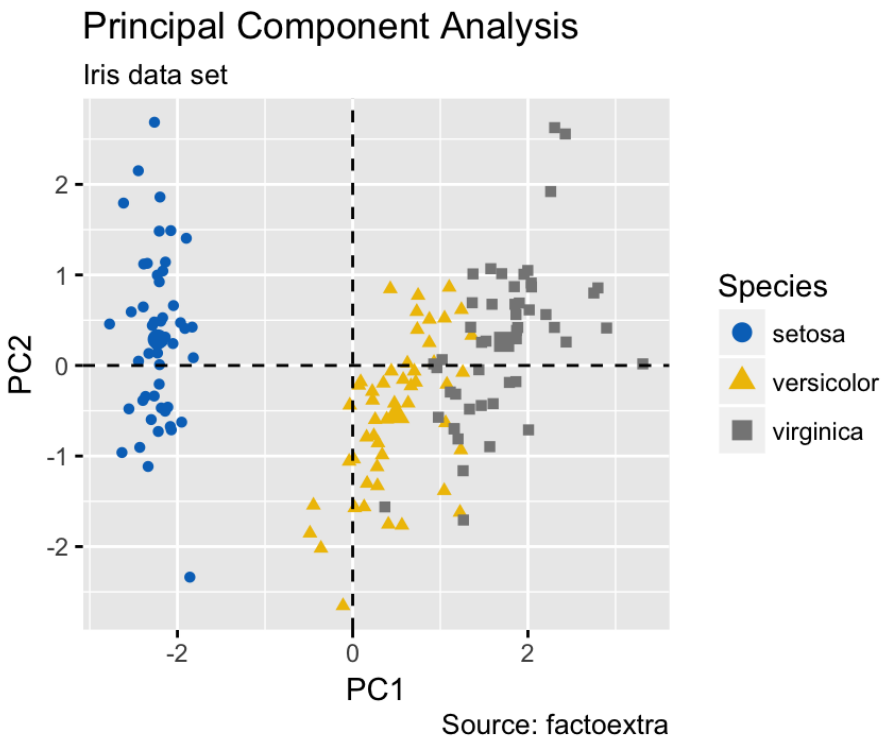

Figure 2: Représentation d'échantillons d'individus sur 2 composantes

4. Calculez deux à deux les corrélations entre les différentes variables du jeu de données (en anglais : pairwise correlations). Représentez ces corrélations sous la forme d'un corrélogramme de votre choix (par ex., avec des bulles et des couleurs). L'objectif est de vous montrer la difficulté à représenter de cette manière l'intégralité des relations entre variables.

5. Représentez maintenant avec Iris un cercle des corrélations, comme en Figure 3. Quelles sont les variables qui semble les plus corrélées, d'après ce que vous avez obtenu à l'étape précédente, et comment cette forte corrélation se matérialise-t-elle dans le cercle des corrélations ?

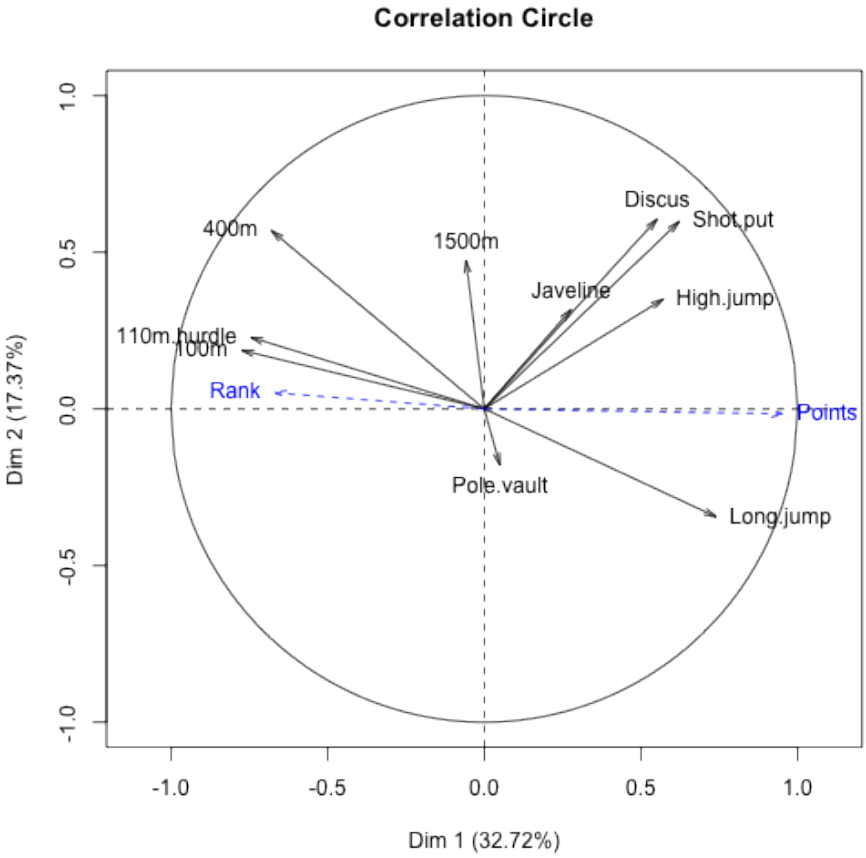

Figure 3: Cercle des corrélations d'une ACP

6. Représentez cette fois le biplot, comme en Figure 4, avec à la fois les variables, et les individus. Quelle part de la variance est-elle représentée sur le plan représenté ?

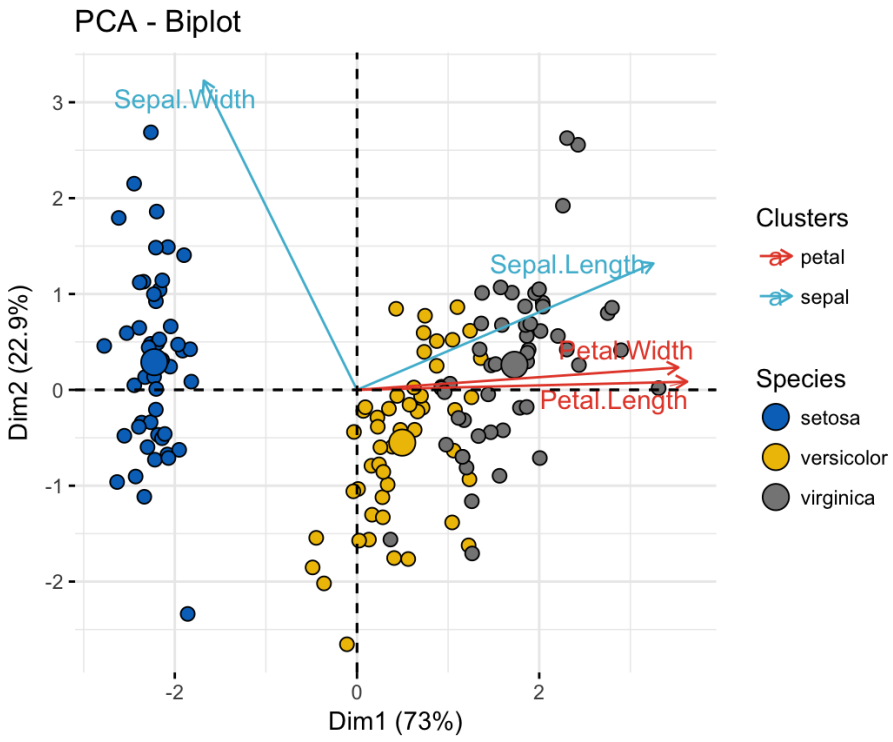

Figure 4: Un biplot, avec les variables et les individus

7. Comme dans la Figure 5, effectuez le test des bâtons brisés à travers un "scree plot", afin de déterminer combien de dimensions il vous paraît le plus pertinent de retenir

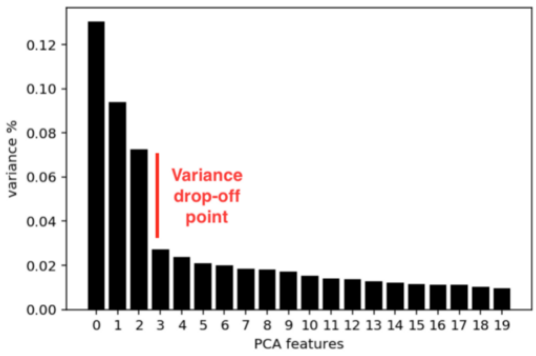

Figure 5: "Scree plot" et méthode du coude

8. Rajoutez sur la figure précédente une courbe permettant de déterminer la part de variance expliquée au fil des variables

9. Produisez une table représentant pour chaque composantes les eigenvalues, et le % de variance expliquée, comme dans la Table 1 ci-dessous. Quel lien semble unir les eigenvalues et le % de variance expliquée ?

Component	Initial Eigen- values	% Variance	Cumulated Variance
1	3.501	58.347	58.347
2	0.853	14.215	72.562
3	0.543	9.051	81.613
4	0.407	6.778	88.390
5	0.371	6.188	94.578
6	0.325	5.422	100.000

Figure 6: Représentation d'échantillons d'individus sur 2 composantes

10. Représentez maintenant à travers une table la saturation (factor loading) des différentes variables sur les composantes principales. Quel est pour la CP1, la variable qui présente la saturation la plus forte ? Quel lien pouvez-vous effectuer avec le cercle des corrélations vu précédemment ?

11. Réalisez une table où pour chaque CP, en ligne, vous représentez la variable dont la saturation est la plus forte pour une CP donnée. Il existe une commande dédiée en Python

	Factor loadings based on correlations matrix Factors			
Character	1	2	3	4
Pla-L	0.26	-0.41	0.41	0.48
Set-L	0.74	-0.13	0.27	0.16
Cap-L	0.79	-0.07	0.18	0.13
Cap-W	0.50	0.15	-0.32	0.44
Ope-L	0.77	0.38	-0.21	-0.26
Pme-L	0.56	0.34	-0.09	0.11
Pte-L	0.79	0.25	-0.18	-0.27
HP-L	0.32	-0.09	0.68	-0.44
PRO-L	-0.26	0.88	0.28	0.12
PRO-W	-0.28	0.87	0.28	0.16

Figure 7: Saturation des variables sur différentes composantes principales

12. Reprenez maintenant le cercle des corrélations fait précédemment. Quel lien effectuez-vous entre saturation des variables et l'orientation des flèches représentant des vecteurs pour chaque variable ?

13. Expliquez ce que signifie, au juste, la qualité de la représentation d'une variable par une ACP. Représentez maintenant le cercle des corrélations, mais la couleur de la flèche doit dépendre la qualité de la représentation des variables par l'ACP. Vous utiliserez les cos2, puis les contributions. Quelle est la principale différence entre ces deux métriques ?

14. Répétez l'étape précédentes, mais en représentant cette fois uniquement les individus sur le plan factoriel. De la même manière, la couleur des individus devra dépendre de la qualité de leur représentation par l'ACP

15. Produisez une table représentant la contribution des individus aux deux premiers axes de l'ACP. Que représente au juste cette contribution ?

16. Pour les utilisateurs de R uniquement : représentez maintenant les individus sur le plan factoriel, mais rajoutez des ellipses. Expliquez en quoi ces ellipses se distinguent d'une classification par un algorithme de type kmeans par exemple.

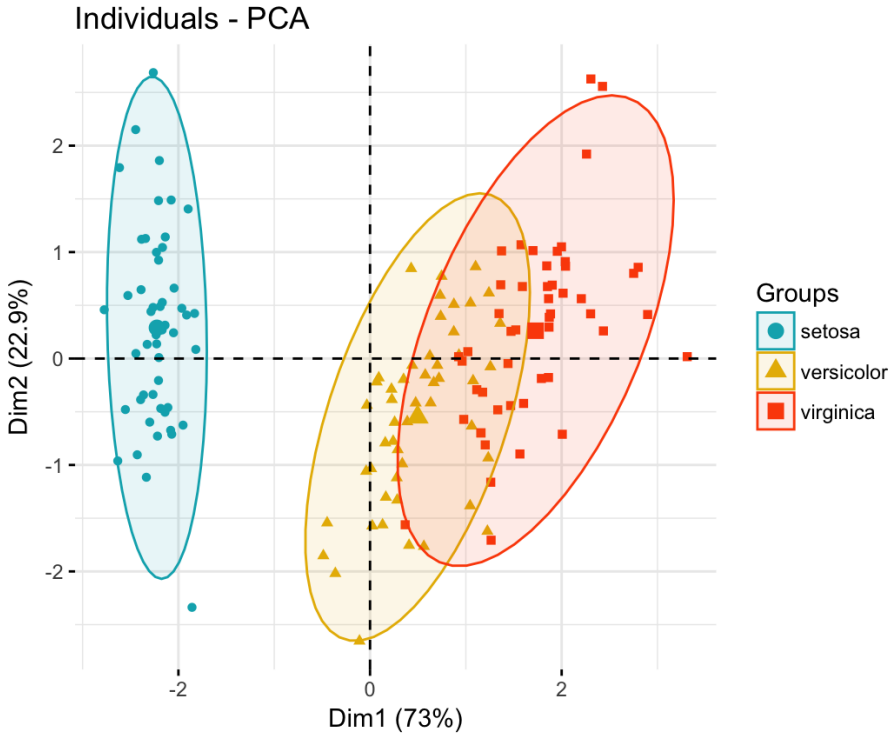

Figure 8: Mise en évidence d'ellipses après une ACP

17. Pour les utilisateurs de R uniquement : réalisez à nouveau un biplot, mais cette fois en considérant la longueur des pétales comme une variable supplémentaire (elle devra apparaître en bleu). Quels sont les principaux changements constatés dans le graphique créé, par rapport à sa version précédente ?

18. Rappelez le fonctionnement de l'algorithme k-means, en insistant sur la différence entre l'étape dite d'expectation, et celle de maximisation. Décrivez la trajectoire que suivent, avec cet algorithme, les différents centroïdes au fil des itérations

19. Appliquez l'algorithme de kmeans pour réaliser un clustering des différentes fleurs recensées dans le jeu de données, puis affichez dans le même plan que celui de l'ACP les différents individus, avec une couleur par cluster. Vous devrez donc prendre en entrée un nombre défini de composantes principales, et non les variables originelles du jeu de données. Vous commencerez avec k=3. Décrivez les caractéristiques de ces trois clusters, en termes de moyennes de variables quantitatives, et de proportions des différentes espèces représentées

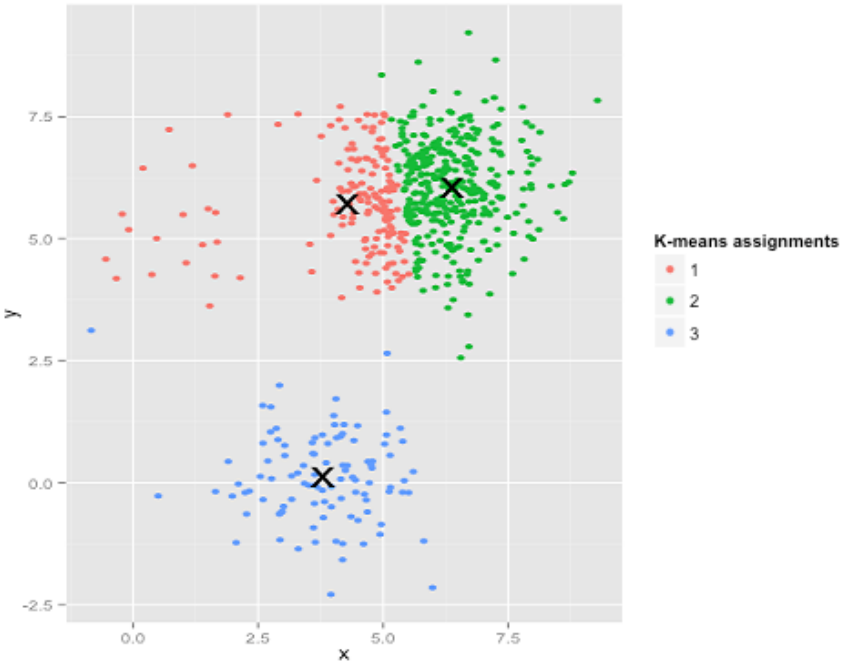

Figure 9: Clustering par k-means, avec centroïdes visibles

20. Explorez grâce à un scree plot l'évolution de la variance intracluster, et appliquez différentes techniques pour déterminer le nombre optimal de clusters à retenir (bâtons brisés, silhouette, nClust pour les utilisateurs de R). Expliquez la logique sous-jacente aux différentes techniques mobilisées.

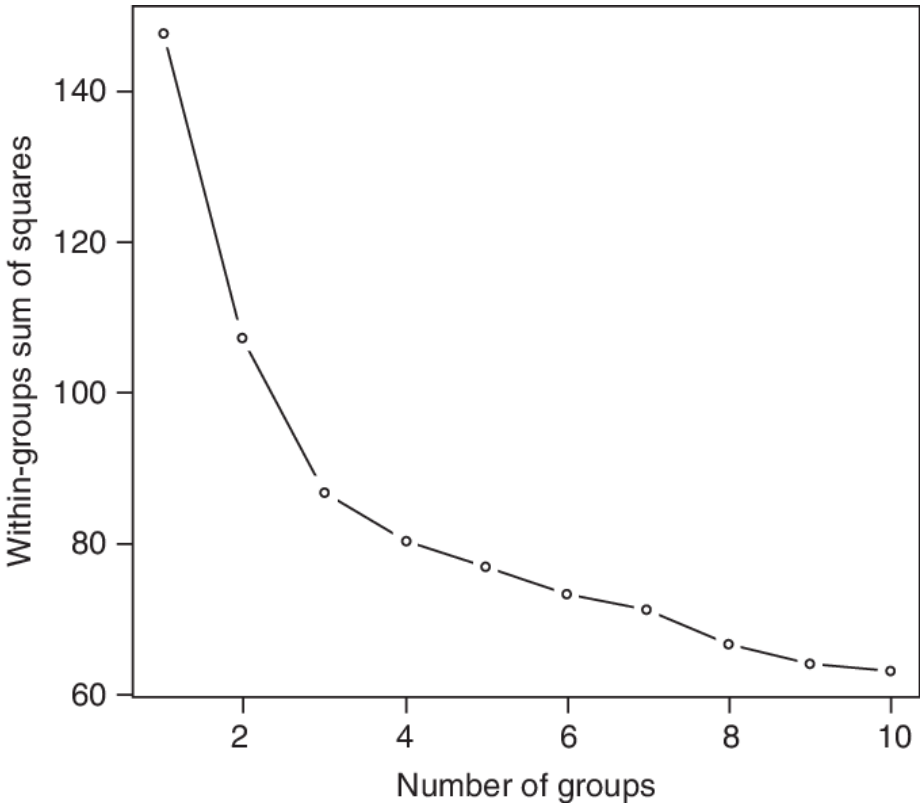

Figure 10: Détermination du nombre optimal de clusters, méthode du coude

3 Projet 2 : S'entraîner sur les données de décathlon

1. Mobilisez les données du jeu de données sur le décathlon, effectuez quelques analyses de réduction de dimensionnalité par ACP, et produisez un cercle des corrélations

2. Dans une seconde figure, représentez les individus sur un plan factoriel. La taille des bulles représentant les individus doit être liée à la qualité de leur représentation

3. Pour les utilisateurs de R uniquement : représentez les individus dans une figure en 3D, chaque axe correspondant à une composante principale, selon le principe illustré ci-dessous

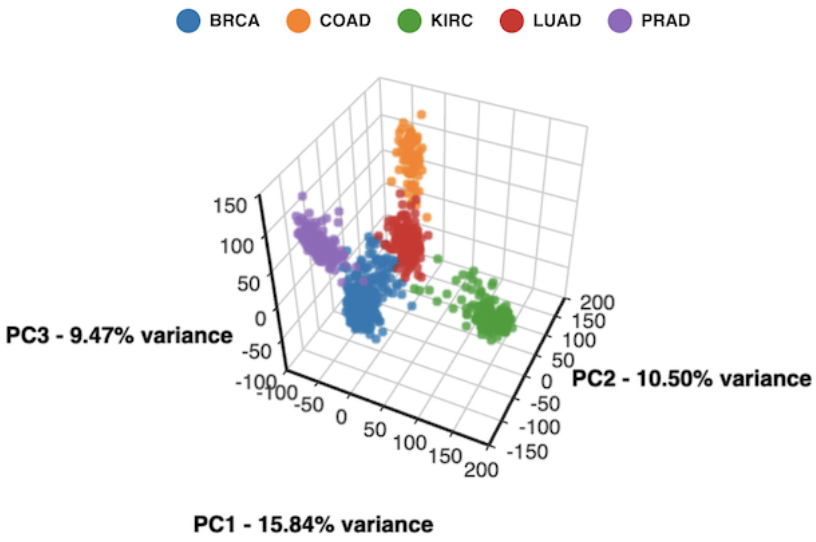

Figure 11: Visualisation d'une ACP en trois dimensions

4. Nommez les cinq premiers axes et justifiez votre choix sur la base d'une table, créée vers la fin du Projet 1

5. Précisez les avantages et inconvénients respectifs des méthodes kmeans et CAH

6. Réalisez un clustering par kmeans, puis un par CAH, et représentez les clusters ainsi créés sur le plan factoriel. Une figure par méthode

7. Dans le cas de la CAH, représentez un dendogramme à plat, comme ci-dessous. Comment avez-vous choisi en définitive le nombre de clusters que vous retenez ?

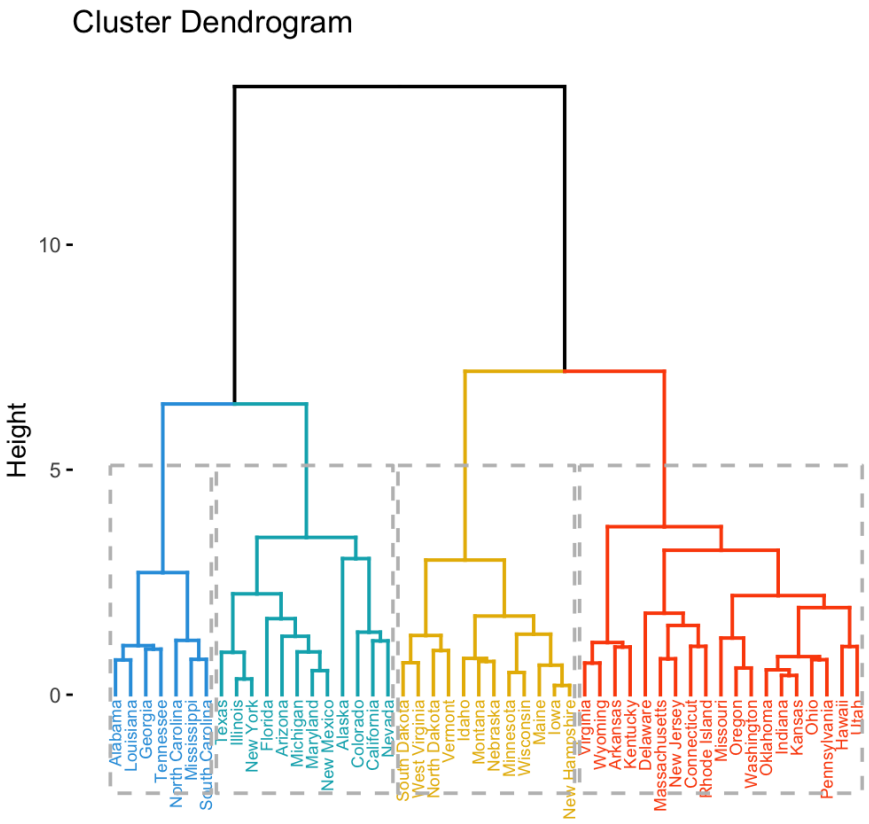

Figure 12: Visualisation d'un dendrogramme

8. Représentez un dendogramme en trois dimensions, avec le plan factoriel, comme ci-dessous

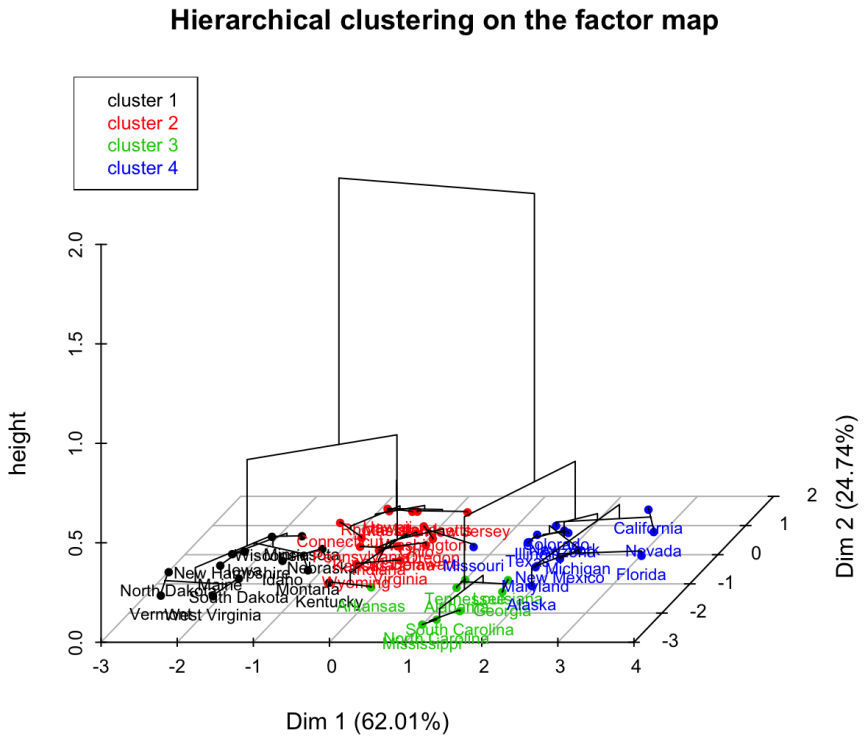

Figure 13: Dendrogramme en 3D et plan factoriel

4 Projet 3 (optionnel) : Dating et analyse des correspondances multiples

Il s'agit ici d'un jeu de données portant sur plusieurs dizaines de milliers d'utilisateurs d'une application de dating imaginaire. Des précisions quant à la signification des variables sont données en annexe.

1. Expliquez dans quelles circonstances l'on mobilise une analyse des correspondances, ou une analyse des correspondances multiples
2. Représentez les différentes variables qualitatives constitutives de ce jeu de données dans un plan factoriel, comme dans la figure ci-dessous. Cantonnez-vous aux variables, ne représentées pas les individus.
3. Que constatez vous quant à l'inertie représentée sur ce plan ?

4. Nonobstant la valeur de la variance associée aux deux premiers axes, tâchez de décrire sans valeur chiffrée ce que vous visualisez dans la dernière figure, en termes de phénomène macrosocial. Il s'agit ici de faire émerger des formes de profils-type d'utilisateurs d'application de dating

5. Représentez contribution des variables sur les deux premiers axes via une table

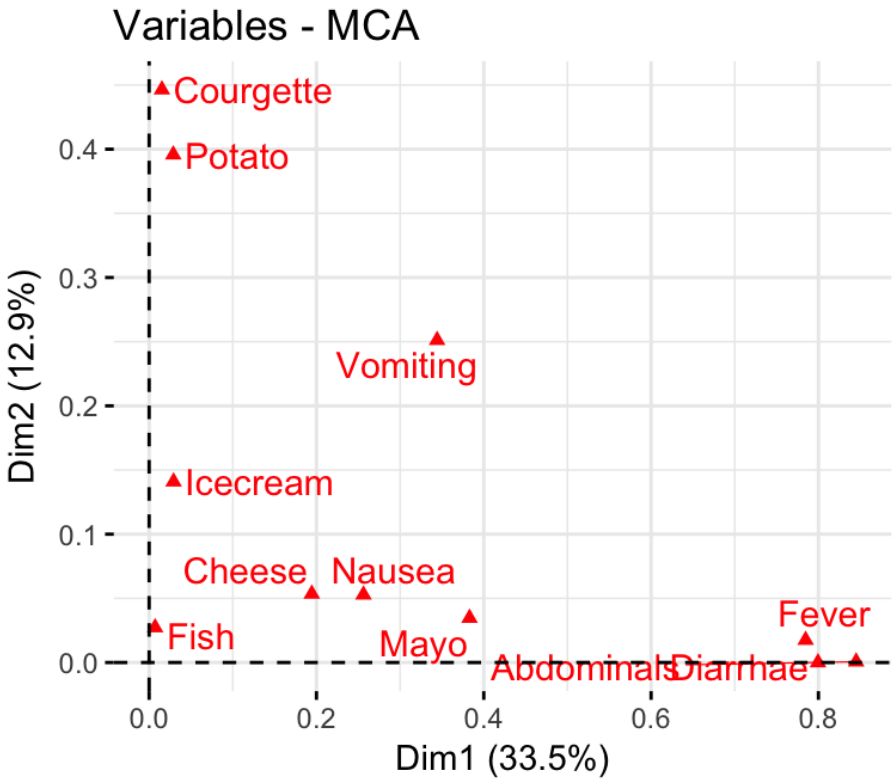

Figure 14: Variables, ACM et plan factoriel

5 Projet 4 (optionnel) : Classification non supervisée par DBSCAN

1. Produisez un jeu de données aléatoires en forme de lunes se faisant face, avec la fonction make moon de Python. Utilisez le jeu de données fourni par l'enseignant si vous utilisez R

2. Utilisez les trois techniques de clustering suivantes pour : k-means, CAH, et DBSCAN. Reproduisez les graphiques correspondants