Transcription de l'audio de la vidéo sur la classification

Nous allons voir comment réaliser une classification avec FactoMineR et Factoshiny. Dans FactoMineR, la classification est réalisée soit sur un tableau de données brut, soit sur un objet résultat d'une analyse factorielle, que ce soit une ACP, une analyse des correspondances, une analyse des correspondances multiples, une analyse factorielle de données mixtes ou une analyse factorielle multiple.

Quand on a un tableau de données avec des variables quantitatives, on peut donc soit réaliser directement la classification sur le tableau, soit faire au préalable une ACP et utiliser les coordonnées des individus de l'ACP pour faire la classification, en conservant uniquement les premières dimensions de l'ACP. Quand les valeurs sont qualitatives, mixtes, ou quand il s'agit d'un tableau de contingence, on utilisera une ACM ou une analyse de données mixte ou encore une analyse des correspondances. La classification sera alors lancée depuis le menu Factoshiny de la méthode factorielle, après être sorti de la méthode d'analyse factorielle.

Nous allons lancer la classification sur le jeu de données décathlon. Petit rappel sur ce jeu de données. Nous avons 41 athlètes et 13 variables. Les 10 premières variables concernent les performances des 41 athlètes aux 10 épreuves du décathlon; donc, par exemple, 100 m, longueur, poids etc. Ensuite on trouve 2 variables quantitatives, utilisées comme variables supplémentaires, le rang de l'athlète lors de l'épreuve et le nombre de points obtenus, et enfin une dernière variable qualitative qui est le nom de la compétition (décastar ou Jeux Olympiques). On réalise la classification à partir des 10 premières variables uniquement, ainsi les distances entre individus sont calculées à partir de leurs performances aux épreuves, sans les variables nombre de points, rang et compétition.

Ouvrons Rstudio et chargeons le package Factoshiny puis le jeu de données decathlon qui est disponible dans FactoMineR. En faisant Factoshiny(decathlon), nous lançons la fonction Factoshiny sur le jeu decathlon, ce qui ouvre l'interface Factoshiny

Si on clique sur classification, la classification est lancée à partir des variables quantitatives brutes directement. Si on fait préalablement une ACP, cela permet de choisir les variables qui servent à calculer les distances entre individus, on peut également normer les variables, ce qui modifie le calcul des distances entre individus en donnant la même importance à chaque variable dans ce calcul, et enfin, on va choisir le nombre de dimensions que l'on souhaite conserver en faisant la classification.

Ici, on va faire une ACP et normer les variables pour donner la même importance à chaque variable dans le calcul des distances. Et on va éliminer les dernières dimensions en conservant seulement les premières. On va donc mettre en supplémentaire les variables nombre de points et rang, ce qui revient à construire la classification à partir de toutes les autres variables quantitatives. On norme les variables, ce qui est fait par défaut, et enfin, pour accorder la même importance à chaque variable, puis on choisit le nombre de composantes à conserver pour la classification. En effet les distances utilisées pour construire la classification seront calculées uniquement à partir des premiers axes factoriels. Les derniers axes étant laissés de côté car considérés comme du bruit. Il y a donc un choix sur le nombre de dimensions qu'il est important de réaliser. En cliquant sur Valeurs, on va voir le tableau des pourcentage d'inertie. Ici on va conserver 8 dimensions car on voit qu'avec 8 dimensions

on conserve 9p6% de l'information. On pourrait se contenter d'un peu moins de dimensions et conserver 75% de l'information par exemple, ou choisir de conserver toute l'information et garder toutes les dimensions.

Après avoir cliqué sur « quitter l'application », l'interface d'ACP se ferme et une nouvelle interface s'ouvre. On peut maintenant paramétrer la classification. Dans un premier temps, on spécifie le nombre de classes qui sera utilisé par la suite pour créer une partition. Un nombre de classes est suggéré. Ce nombre maximise la différence de perte d'inertie entre deux partitions successives. Autrement dit, on retiendra la partition en Q classes si la perte est importante entre une partition en Q-1 classes et une partition en Q classes, et que cette perte est faible entre la partition en Q classes et celle en Q+1 classes. Nous allons garder le nombre par défaut qui est de 4 classes.

On peut choisir de faire un prétraitement par K-means avant de faire la classification. Ceci est utile quand il y a beaucoup d'éléments à classer. L'idée est alors de construire une partition grossière avec beaucoup de classes, par exemple une centaine, et de construire ensuite l'arbre hiérarchique à partir des 100 centres de classes (pondérés par l'effectif de la classe). Le haut de l'arbre hiérarchique est stable par rapport à une classification construite à partir de tous les individus, mais la classification va être beaucoup plus rapide. Et surtout elle devient possible s'il y a vraiment beaucoup d'individus alors que ce n'est pas le cas si on construit l'arbre à partir de tous les individus. Attention toutefois, s'il y a vraiment beaucoup d'individus, plusieurs centaines de milliers, on conseille de travailler sans l'interface shiny et de lancer les lignes de code suivantes : une première pour faire l'ACP ou toute autre analyse factorielle, la suivante pour faire la classification sans décrire les classes en utilisant l'argument kk pour faire un prétraitement kmeans avec kk =100 classes. Ensuite on peut construire les graphes puis caractériser les classes.

```
res.pca <- PCA(MyData, ncp=8, graph=FALSE) ## ACP en conservant 8 dimensions
hc <- HCPC(res.pca, kk=100, description=FALSE, graph=FALSE) ## classification avec
prétraitement par kmeans avec kk=100 classes
plot(hc,choice = "tree")                    ## graphe de l'arbre
plot(hc,choice = "map", draw.tree=FALSE)    ## plan d'ACP avec les classes
plot(hc,choice = "3D.map")                  ## plan d'ACP avec arbre en 3D
catdes(hc$data.clust, ncol(hc$data.clust))  ## caractérisation des classes
```

Consolidation permet d'améliorer la partition obtenue en coupant simplement l'arbre hiérarchique. Sans consolidation, les classes sont celle obtenues en coupant simplement l'arbre. Si on consolide les classes, alors les individus sont réaffectés au centre de classes le plus proche par un algorithme de K-means. Les classes seront plus homogènes mais on perd la correspondance entre les classes et l'arbre hiérarchique, ce qui est un inconvénient. Nous allons consolider les classes.

Enfin on peut choisir la distance Euclidienne ou la distance de Manhattan. Si on travaille sur les résultats d'une analyse factorielle, on utilisera toujours la distance euclidienne. Si on travaille sur les données brutes, on travaille généralement avec la distance euclidienne mais dans certains cas la distance de Manhattan est utile.

Maintenant que nous avons paramétré la classification, nous pouvons commenter les résultats. Il y a quelques options graphiques qui permettent de modifier les graphes. Par exemple, dans la représentation 2D, on peut projeter l'arbre sur le graphe. Ou bien dans le graphe 3D, on peut ajouter les noms des individus ou encore représenter les centres de classes.

Dans le premier graphe, on a un diagramme avec les gains d'inertie. On voit ici une première barre qui indique qu'il y a une forte perte d'inertie si on passe de 2 classes en 1 classe, donc il ne faut pas regrouper 2 classes pour en faire 1 seule; même chose, il y a une perte d'inertie si on passe de 3 à 2 classes et si on passe de 4 à 3 classes. Par contre, en passant de 5 à 4 classes, la perte d'inertie est beaucoup plus faible. On peut donc garder 4 classes ici. Un autre critère pour choisir le nombre de classes et la forme de l'arbre. Ici l'allure de l'arbre suggère que le niveau de coupure proposé convient et on garde le découpage en 4 classes.

Le second graphe correspond aux individus sur le plan principal de l'ACP, donc sur les dimensions 1 et 2. Les individus sont coloriés en fonction de l'appartenance à leur classe. Il y également un graphe en 3 dimensions avec les individus qui sont positionnés sur le plan 1-2 de l'ACP avec, en perspective, l'arbre hiérarchique. Cet arbre hiérarchique montre les proximités entre les individus. On voit par exemple ici, comme sur le graphe précédent, que les classes bleue et noire sont bien séparées mais que les classes verte et rouge sont beaucoup moins bien séparées. On peut se poser la question: pourquoi ces classes sont enchevêtrées? On peut construire le graphe sur les dimensions 3 et 4 et voir que les classes rouge et verte sont cette fois bien séparées. Donc en fait, le premier plan sépare bien les classes bleue et noire, tandis que le plan 3 – 4 sépare bien les classes rouge et verte. Donc on a besoin ici de 4 dimensions pour séparer les 4 classes.

On peut maintenant voir les résultats numériques. Les onglets « Caractérisation des classes » et « Lien avec la partition » permettent de visualiser les tableaux, tandis que l'onglet « Valeurs » détaille les résultats. Si on clique sur « Caractérisation des classes », on a en colonne chaque classe et en ligne les variables quantitatives et les modalités des variables qualitatives. Les variables ou modalités présentent caractérisent au moins une classe au seuil choisi (par défaut 0.05). Les lignes sont triées par ordre alphabétique, mais on peut cliquer sur une classe pour trier les variables par rapport à leur pouvoir discriminant de la classe. Par exemple, en cliquant sur classe 1, on voit que la variable nombre de points est très caractéristique. La valeur du tableau est une valeur-test. Inférieure à -2, cela signifie que les individus de cette classe obtiennent un nombre de points significativement inférieur au nombre de points obtenu par un individu en général. Ici aucune modalité ne caractérise les classes mais si on augmente la probabilité en mettant une valeur de 0.5, alors on voit apparaître les 2 modalités décastar et JO.

L'onglet « lien avec la partition » donne les variables qui permettent de séparer au mieux les classes, i.e. qui permettent de caractériser la partition. Un tableau est fourni avec les variables quantitatives et un autre pour les variables qualitatives. Par exemple ici, on voit que c'est la variable Points qui permet de séparer au mieux les classes.

L'onglet « Valeurs » détaille ces résultats obtenus dans les 2 onglets précédents. Et permet aussi de décrire les classes en fonction des individus. On a donc plusieurs objets de résultats : une description des classes par les variables, une description par les axes, et une description par les individus.

La description par les variables décrit d'abord les classes par les variables quantitatives (on retrouve les infos qui sont dans l'onglet « liens avec la partition »). Les variables sont triées de celles qui permettent de décrire le mieux la partition à celles qui permettent de décrire un peu moins bien la partition mais qui permettent de décrire quand même de façon significative. Seules les variables qui ont une liaison significative avec la variable de classe sont conservées ici. L'intensité de la liaison est mesurée par le rapport de corrélation entre la variable quantitative et la variable de classe. On regarde

si ce rapport de corrélation est significativement différent de 0. On voit par exemple ici que la variable nombre de points est la variable la plus liée à la variable de classe.

Ensuite on retrouve la description de chacune des classes avec des résultats plus détaillés. Par exemple, pour la première classe, les variables les plus caractéristiques de la classe sont le 100m, le 400m, mais également le nombre de points et le poids. Pour le 100m, les individus de la classe 1 prennent des valeurs significativement différentes de 0 et supérieures à la moyenne. Significativement différentes de 0 car la valeur-test est supérieure à 2 en valeur absolue, et supérieures à la moyenne parce que la valeur-test est positive. Dans le détail, les individus de la classe 1 courent le 100m en 11''27 en moyenne alors que les individus de toutes les classes, y compris de la classe 1, courent le 100m en 10''99. Les individus de la classe 1 ne courent pas très vite puisqu'ils mettent plus de temps pour courir le 100m. A partir du nombre de points, on peut dire que ce sont des individus qui ont un nombre de points significativement plus faible que les autres (plus faible parce que la valeur-test est négative). En moyenne les individus de la classe 1 ont obtenu 7596 points quand la moyenne générale sur l'ensemble des athlètes est de 8500 points. On a ensuite une description de la classe 2, de la classe 3 et de la classe 4. Ici, aucune variable qualitative ne permet de décrire les classes car sinon on aurait une description des classes par les variables qualitatives également.

On peut caractériser les classes par les axes, donc les dimensions factorielles de l'ACP dans notre cas. Les dimensions factorielles étant des variables quantitatives, on utilise la même méthodologie que précédemment et on voit que les dimensions 1 et 3 permettent de caractériser au mieux les classes. En effet la classe 1 a des coordonnées significativement plus faibles que les autres sur la dimension 1; pour les individus de la classe 2, ils ont des coordonnées significativement plus faibles sur la dimension 3; les individus de la classe 3 ont eux des coordonnées significativement plus fortes sur la dimension 3 et un peu plus faibles sur la dimension 2; et les individus de la classe 4 ont des coordonnées significativement plus fortes sur la dimension 1. Quand on voit ça, on peut représenter l'arbre hiérarchique sur le plan 1-3, ce que l'on fait ici. Effectivement avec une représentation de l'arbre hiérarchique sur le plan 1-3, on voit bien la séparation des classes.

Une petite remarque sur la représentation avec le plan 1-2 : cette représentation permet d'avoir la meilleure vision des distances entre les individus. Mais en ajoutant les couleurs sur les différentes classes, on voit qu'il y a une différenciation des individus de la classe verte et de la classe rouge, donc une forte distance entre les individus de la classe verte et la classe rouge. La classification permet de voir en plus ce qui se passe sur les dimensions 3 et 4. Donc on a une vision optimale des distances par le plan d'ACP complétée par la classification.

Nous pouvons voir maintenant les résultats sur les individus et dans un premier temps trouver les parangons de chaque classe, c'est-à-dire les individus les plus proches du centre de la classe. Pour la classe 1, le parangon est Uldal qui est à une distance de 1.43 par rapport au barycentre des individus de la classe 1. C'est le plus proche du centre la classe. Puis le 2ème plus proche est Barras ensuite Karlivans, etc. Pour les individus de la classe 2, le plus proche est Hernu, etc.

Nous avons une autre mesure à partir des individus, ce qu'on appelle la spécificité. Pour ce faire, on calcule la distance des individus au barycentre des autres classes. Pour les individus de la classe 1, Casarsa est l'individu qui est le plus loin des autres classes, plus loin des barycentres des autres classes. La distance au barycentre le plus proche est de 5.08. De ce point de vue, Casarsa est vraiment très

spécifique de la classe 1, il ne pourrait pas être dans les autres classes. L'individu le plus spécifique de la classe 2 est Smith puisqu'il est très loin des barycentres des autres classes.

L'interface propose un rapport automatique sur les résultats de la classification en anglais ou en français. Ce rapport automatique récupère les graphes et la caractérisation des classes et peut être généré au format Rmarkdown, html ou word.

Enfin, le bouton « lignes de codes » de la classification donne les lignes de codes permettant de mettre en œuvre la méthode à l'identique. En cliquant sur lignes de codes, les lignes de code apparaissent pour mettre en œuvre l'analyse factorielle (ACP, ACM ou autre), puis la classification et enfin pour dessiner les graphes.

Vous avez vu tous les résultats sur la classification. On peut décrire donc les classes à partir des individus, à partir des variables, quantitatives, qualitatives, actives ou supplémentaires, à partir des dimensions factorielles. Et puis nous avons les résultats aussi sur l'arbre hiérarchique. A vous de jouer maintenant pour mettre en œuvre des classifications.