Transcription de l'audio de la vidéo sur l'ACP

Nous allons voir comment mettre en œuvre l'ACP avec FactoMineR. Pour ce faire, nous allons utiliser un jeu de données qui s'appelle décathlon. Le jeu de données a la forme suivante: il y a 41 lignes qui correspondent à 41 athlètes et 13 variables. Parmi ces 13 variables, nous avons 10 variables qui concernent les performances aux différentes épreuves du décathlon. Par exemple, 100m, longueur, poids, hauteur, 400m, etc. Ensuite 2 variables quantitatives que sont le rang de l'athlète à la compétition et le nombre de points. En effet, au décathlon, chaque épreuve apporte un nombre de points à l'athlète et donc l'athlète accumule un certain nombre de points en fonction des 10 épreuves. On a le total de points de l'athlète lors de la compétition. Il y a également une variable compétition avec 2 modalités: la modalité decastar et la modalité jeux olympiques 2004. C'étaient 2 grandes compétitions qui ont eu lieu en 2004.

Nous allons réaliser l'ACP sur ce jeu de données en considérant les 10 épreuves comme les variables actives, les variables rang et nombre de points seront considérées comme supplémentaires ainsi que la variable compétition. De ce fait, nous allons comparer les athlètes du point de vue uniquement de leurs performances aux 10 épreuves.

Passons maintenant sur R pour mettre en œuvre l'ACP. Ouvrons donc Rstudio ou R. Nous allons commencer par importer le jeu de données. Le jeu de données est disponible via le lien suivant. Nous allons donc l'importer en utilisant la fonction read.table, en précisant que le nom des variables est disponible dans le jeu de données, que le séparateur de colonnes est le ";", ce qui est le plus fréquent lorsqu'on utilise des fichiers csv, que le nom des individus est disponible dans la première colonne du jeu de données avec row.names=1, fileEncoding=latin1 permet d'éviter les problèmes d'encodage des accents avec les mac, et enfin je précise qu'il ne faut pas vérifier le nom des variables en mettant check.names=FALSE. Cela signifie que le nom des variables sera pris tel quel dans le jeu de données.

Le jeu de données est importé dans un objet decathlon. Vérifions que l'importation a été faite correctement. Nous voyons que les variables quantitatives 100m, longueur, poids, etc. jusqu'au classement et au nombre de points sont bien quantitatives puisqu'on voit un minimum, une moyenne, un maximum, et les quartiles. Et enfin, la dernière variable, la variable compétition, est bien qualitative, on voit le nombre de fois où la modalité decastar a été choisie, 13 fois, et la modalité jeux olympiques choisie, 28 fois.

Voyons maintenant comment lancer l'ACP avec FactoMineR, et plus précisément Factoshiny son interface graphique. Cette interface lance les commandes de FactoMineR et il n'est pas nécessaire de connaître la syntaxe de R. Cette interface permet aussi d'améliorer la lisibilité des graphiques. Chargeons le package Factoshiny. Pour lancer l'ACP, comme toute autre méthode d'analyse de données disponible dans l'interface, il suffit d'utiliser la fonction Factoshiny sur le jeu de données. Cette fonction peut être lancée sur un jeu de données, sur un objet résultats d'ACP ou sur un objet résultat de la fonction Factoshiny. Lançons la fonction sur le jeu de données decathlon.

L'interface graphique s'ouvre dans le navigateur par défaut. La fenêtre est divisée en 2 parties. Sur la gauche, on trouve un descriptif succinct du jeu de données, puis les méthodes qui peuvent être

appliquée sur ce jeu de données, et un lien vers une vidéo qui aide au choix de la méthode à utiliser. Sur la partie de droite, on trouve les différentes méthodes. En cliquant sur l'aide d'une méthode, on trouve une description rapide de la méthode ainsi que des liens vers des vidéos de cours sur la méthode. Si on clique ensuite sur « lancer », l'analyse est exécutée et une nouvelle fenêtre s'ouvre dans le navigateur. Cette nouvelle fenêtre est divisée en 2 parties. Sur la gauche, on trouve le menu qui va permettre de paramétrer la méthode ou les graphes, sur la droite on trouve les résultats. Dans le menu de gauche, nous avons plusieurs onglets. Le premier va servir à paramétrer la méthode, i.e. à choisir les variables qui seront actives et les variables qui seront illustratives, les individus actifs ou supplémentaires et également la gestion des données manquantes si des données manquantes sont présentes dans le jeu de données.

Nous avons ensuite un onglet qui permet d'améliorer les graphiques, un onglet qui permet de réaliser une classification à l'issue de l'ACP. Un onglet pour obtenir un rapport automatique qui permet d'interpréter les principaux résultats de l'ACP. Et enfin 2 boutons. Un bouton qui va permettre de récupérer les lignes de code de l'ACP pour pouvoir remettre en œuvre l'ACP et reconstruire les graphiques tels qu'ils sont dans l'interface et un bouton pour quitter l'application.

Dans un premier temps, paramétrons la méthode. Nous allons cliquer sur l'onglet « paramètres de l'ACP ». Je vais donc choisir les variables quantitatives qui seront supplémentaires. Les autres variables quantitatives seront actives. Dans l'exemple, les variables nombre de points et classement sont quantitatives supplémentaires. La variable compétition est qualitative supplémentaire. Par défaut, toutes les variables qualitatives seront utilisées comme variables supplémentaires. Si on ne veut pas que certaines variables soient utilisées comme supplémentaires, il faut les supprimer. Et je n'ai pas d'individus supplémentaires, donc je vais laisser cette case vide. Par défaut, les variables sont centrées-réduites. Si on ne voulait pas réduire les variables, il faudrait décocher cette case. Et enfin, il y a la possibilité de gérer les données manquantes. Dans cet exemple, il n'y a pas de données manquantes. Si le jeu de données contenait des données manquantes, nous aurions ici la possibilité de choisir différentes méthodes d'imputation. Voici ce que nous aurions. La possibilité d'imputer par la moyenne de la variable. Cette méthode est très rapide mais pas recommandée. La possibilité d'imputer par un modèle d'ACP à 2 dimensions, ce qui est plutôt un bon compromis dans la plupart des situations. Et enfin la possibilité d'imputer par un modèle d'ACP à k dimensions. Le nombre k est dans un premier temps estimé par validation croisée, ce qui peut être un peu long sur de gros jeux de données. Mais le nombre de dimensions est alors optimal pour le modèle d'ACP qui servira à imputer les données manquantes du jeu de données. Une fois qu'on a imputé le jeu de données, on se retrouve avec un jeu de données complet sur lequel on va pouvoir mettre en œuvre l'ACP.

Revenons à notre jeu de données qui n'a pas de données manquantes. Nous avons fini de paramétrer la méthode. Nous allons soumettre pour pouvoir prendre en compte toutes les modifications que nous avons faites, soit en cliquant sur ce bouton, soit en cliquant sur cette case pour sortir de l'onglet paramètres. Ainsi, les nouveaux paramétrages de la méthode sont pris en compte et les résultats de l'ACP sont mis à jour avec les deux variables classement et nombre de points supplémentaires et la variable compétition comme variable qualitative supplémentaire.

Nous avons sur la droite les principaux résultats regroupés dans cinq onglets : un onglet sur les graphes, un onglet avec les principaux résultats quantitatifs, un onglet sur la description automatique des dimensions et puis un résumé du jeu de données et le tableau de données.

Voyons dans un premier temps les résumés des principaux résultats. Nous avons ici un listage avec les principaux résultats et dans un premier temps les résultats sur les pourcentages d'inertie associés à chaque dimension. Nous voyons que la première dimension résume 33% de l'information tandis que la deuxième dimension résume environ 17% d'information. Nous avons ensuite les résultats sur les individus, par défaut pour les 10 premiers individus. Si nous voulons les résultats pour l'ensemble des individus, il suffit de mettre le nombre d'individus ici où un nombre plus grand que le nombre d'individus du jeu de données. Si je mets 100, j'aurai les résultats sur tous les individus puisqu'il y a 41 individus. Et donc dans les résultats sur les individus, on a le nom des individus puis la distance de l'individu au centre de gravité du nuage puis les résultats sur la première dimension avec la coordonnée de l'individu sur la 1ère dimension, sa contribution à la construction de la 1ère dimension et sa qualité de représentation sur la première dimension mesurée par le cosinus carré, qui vaut par exemple pour le 1er individu 0.695. Donc on a ce résultat sur la première dimension et on a les mêmes résultats sur la 2ème dimension puis la 3ème dimension. Donc si on voulait voir la qualité de représentation du 1er individu pour le plan, il suffirait de sommer 0.695 et 0.080. Donc voici les résultats pour les individus.

On a ensuite les résultats pour les variables; avec le nom des variables puis, les coordonnées sur la 1ère dimension, la contribution de la variable à la construction de l'axe et la qualité de représentation mesurée par le cosinus carré. Tout ça pour la 1ère dimension puis, la 2ème dimension et la 3ème dimension. Si des variables quantitatives supplémentaires sont présentes dans l'analyse, on a un tableau avec la coordonnée de chaque variable sur la première dimension, sa qualité de représentation, bien sûr on n'a pas sa contribution puisque les variables supplémentaires n'ont pas contribué à la construction des axes. Pour les variables qualitatives supplémentaires, on a les résultats pour chaque modalité de toutes les variables qualitatives, avec la distance au barycentre, la coordonnée sur la première dimension, la qualité de représentation et une v-test qui va mesurer si la coordonnée est significativement différente de 0. La v-test suit une loi Normale et donc les valeurs extrêmes nous indiqueront quelles sont les valeurs qui sont significativement différentes de 0 sur un axe. Et donc on a là encore les résultats sur les dimensions une, 2 puis 3.

Nous avons ensuite un détail de chacun de ces résultats. Par exemple, dans le tableau valeurs propres, avec les résultats sur les pourcentages d'inertie et un graphe avec les pourcentages d'inertie associés à chaque dimension. Et puis même chose, les résultats sur les variables, sur les individus, les variables supplémentaires et les variables qualitatives supplémentaires. On retrouve les mêmes résultats que ceux qui sont résumés dans le premier onglet.

Nous avons ensuite une description automatique des axes, plus précisément des trois premières dimensions, en fonction des variables quantitatives ou qualitatives. Pour les variables quantitatives on a calculé le coefficient de corrélation entre les coordonnées des individus sur l'axe et la variable. Par exemple, la variable nombre de points est corrélée positivement avec la première dimension, la corrélation vaut 0,96 ce qui signifie qu'il y a une corrélation très élevée entre la coordonnée et le nombre de points. Donc les individus qui ont une faible coordonnée sur la première dimension ont obtenu un faible nombre de points. Et cette corrélation est significativement différente de 0 puisqu'on a une probabilité critique inférieure à 5% ici. Sont conservées uniquement les corrélations qui sont significativement différentes de 0. Donc les variables sont triées, des plus liées, de façon significative, au plus liées négativement tout en bas. Je peux changer ici le seuil permettant de conserver les variables, et conserver les variables qui ont une probabilité critique inférieure à 20% plutôt que 5%. On va conserver un petit peu plus de variables. Je fais ça juste pour montrer qu'on peut voir les liaisons

avec des variables quantitatives mais également avec des variables qualitatives. Par exemple, ici, pour la 1ère dimension: on voit que la variable compétition est liée à la 1ère dimension et c'est significatif au seuil de 20%. Donc le R^2 correspond au rapport de corrélation. 5% de la variabilité des coordonnées est expliquée par la variable compétition, ce qui est significatif au seuil 20% (mais pas au seuil classique de 5%). Et puis, au niveau des modalités, si on rentre dans le détail, on voit que les modalités jeux olympiques et decastar sont liées aussi si on utilise le seuil 20%. On utilise ici le seuil 20% uniquement pour un aspect pédagogique, pour montrer qu'on peut caractériser les axes par des variables qualitatives. Mais généralement on utilisera le seuil 5%.

Revenons maintenant sur l'onglet des graphiques. On trouve les graphes par défaut avec, pour le graphe des individus, les individus en noir et les modalités supplémentaires en rose, et pour le graphe des variables, les variables actives en noir et les variables illustratives ou supplémentaires en bleu et en pointillés. Il est souvent très utile de travailler ses graphes pour mieux mettre en évidence les informations. Nous pouvons donc choisir différentes options graphiques ici. Dans un premier temps nous pouvons modifier le choix des axes que nous allons dessiner. Par défaut, ce sont les axes 1 et 2 qui sont dessinés mais nous pouvons par exemple décider de dessiner les axes 3 et 4. Il suffit de modifier ici et choisir 3 et 4 pour dessiner le plan 3-4 pour le graphe des individus et celui des variables. Les deux graphes sont modifiés simultanément puisqu'on les commente en même temps. Remettons les dimensions 1 et 2 et nous pouvons ensuite travailler le graphe des individus ou le graphe des variables.

Commençons par le graphe des individus en modifiant le titre. On peut aussi choisir les points à dessiner et supprimer les modalités supplémentaires pour ne dessiner que les individus actifs. Ce n'est pas très utile dans notre exemple mais quand on a beaucoup de variables qualitatives, cela évite que le graphe soit surchargé par trop de modalités. On peut aussi, si on a beaucoup d'individus, supprimer les libellés des individus et ne conserver que ceux des modalités supplémentaires. On peut augmenter la taille de la police ou encore choisir de ne mettre le libellé des individus que pour certains individus. Par exemple, uniquement pour les individus qui sont bien représentés sur le plan principal, ceux qui ont une qualité de représentation supérieure à 0,5 sur le plan principal seront avec un libellé. On peut aussi sélectionner les individus en fonction de leur contribution à la construction du plan : en choisissant les dix individus qui ont le plus contribué à la construction du plan principal donc des deux premières dimensions. On peut aussi colorier les individus en fonction de la qualité de représentation (du cos2). Ou colorier les individus en fonction d'une variable quantitative comme, par exemple, la variable nombre de points. Si je dessine tous les points, je vais voir que la variable nombre de points est corrélée positivement à la première dimension. On le voit très bien avec le code couleur : nous trouvons à gauche des individus en bleu qui prennent de faible valeur sur le nombre de points, au milieu des individus en violet qui prennent des valeurs moyennes et à droite des individus en rouge qui prennent de fortes valeurs sur le nombre de points. C'est bien la corrélation entre les coordonnées et la variable nombre de points qu'on illustre ici avec les couleurs.

Nous pouvons également colorier les individus en fonction d'une variable qualitative, plus exactement en fonction des modalités qu'ils prennent pour une variable qualitative. Dans notre exemple, il n'y a qu'une variable qualitative donc on va habiller les individus en fonction de la variable competition. On voit ici que les individus coloriés en noir ont participé au decastar et les individus coloriés en rouge ont participé aux jeux olympiques.

On peut aussi construire des ellipses de confiance autour des barycentres decastar et jeux olympiques. Ces ellipses, il faut les voir comme des zones de confiance d'un barycentre. C'est-à-dire que si on avait eu d'autres athlètes qui participaient au decastar, la moyenne se retrouverait, avec un niveau de confiance de 95%, dans cette ellipse. On voit que les zones de confiance de décastar et JO se chevauchent et donc il n'y a pas de différence significative dans la position de ces modalités dans le plan principal.

Nous pouvons maintenant regarder le graphe des variables. Modifier évidemment le titre, augmenter la taille des libellés, et sélectionner les variables en fonction de leur qualité de représentation par exemple pour ne conserver que celles qui ont une qualité de représentation supérieure à 0.6.

On peut éventuellement augmenter la taille des graphiques ici, et bien sûr on peut télécharger les graphiques au format jpeg, png ou pdf.

A l'issue de l'ACP, il est possible de réaliser une classification. Il suffit de cocher cette case ici et de choisir le nombre de dimensions qu'on va vouloir conserver pour construire la classification. Si on conserve uniquement les premières dimensions de l'ACP, cela revient à conserver les dimensions qui contiennent le signal, l'information, et à supprimer les dernières dimensions qui contiennent plutôt du bruit. Ainsi, on aura une classification plus stable. L'idée est donc souvent de conserver les premières dimensions, i.e. celles qui vont permettre de récupérer 70 ou 80% de l'information. Mais on peut également conserver toutes les dimensions de l'ACP, ce qui revient à faire une classification sur les données initiales ou plus exactement les données centrées réduites. Je ne vais pas faire la classification ici car la classification est expliquée dans une autre vidéo.

Il est aussi possible d'obtenir un rapport sur les résultats de l'ACP, i.e. une interprétation automatique simple des résultats de l'ACP en anglais ou en français. Ce rapport automatique peut utiliser des graphes suggérés par l'analyse ou alors utiliser les graphes que l'on vient de travailler. Dans un premier temps, il est intéressant de voir les graphes suggérés par la méthode. On va pouvoir récupérer ce rapport automatique sous différents formats : au format Rmarkdown, au format html ou au format word. Je vais récupérer le rapport au format html. La rédaction du rapport prend un petit peu de temps. Voici donc le rapport qui commence par préciser sur quel jeu de données a été réalisé l'analyse : ici un jeu de données avec 13 variables dont 2 qui sont considérées comme quantitatives supplémentaires et une variable qualitative supplémentaire. Il y a dans un premier temps, une recherche d'individus extrêmes. Les individus extrêmes sont les individus qui contribueraient énormément à la construction des axes et sans qui les dimensions seraient très différentes. Ici il n'y a pas d'individus extrêmes. Ensuite, on commente les pourcentages d'inertie associé à chaque axe. Le graphique donne la décomposition de l'inertie avec le pourcentage d'inertie sur chaque axe et on commente ces pourcentage d'inertie par rapport à ce qui aurait été obtenu pour des jeux de données de même dimension (même nombre d'individus et même nombre de variables actives) si les variables n'étaient pas structurées. Cela permet de voir à quel point les premières dimensions résument l'information. Le commentaire suggère aussi le nombre de dimensions qu'il faudrait interpréter. Ensuite, nous avons une description du plan des individus, avec certains individus qui ont un libellé et d'autres qui n'en ont pas pour éviter d'avoir des graphes trop surchargés. Un graphique avec les individus coloriés selon une variable qualitative. Ici nous avons une seule variable qualitative, mais si plusieurs variables étaient disponibles, le test de Wilks permettrait de choisir la variable pour laquelle les modalités sont le plus séparées sur le plan. Ainsi, les individus seraient coloriés selon cette variable.

On a ensuite le graphe des variables là encore avec certains libellés uniquement pour certaines variables, et puis un graphe avec les modalités supplémentaires. Enfin, il y a une interprétation, ou une tentative d'interprétation, de chacune des dimensions grâce au graphe des individus et des variables. Cette interprétation est très limitée et permet juste de constater les liaisons et les caractéristiques des individus. A chacun ensuite de chercher à aller au-delà de cette interprétation dans l'explication des dimensions notamment. Comme ici 3 dimensions sont proposées, il y a également une description de la troisième dimension. Le rapport automatique a suggéré quelques graphes, il est également possible d'utiliser les graphes que l'on vient de réaliser avec l'interface.

Enfin il y a un bouton « lignes de codes » de l'ACP qui récupère les lignes de codes de l'ACP pour mettre en œuvre la méthode et reconstruire les graphes à l'identique. En cliquant sur lignes de codes de l'ACP, 3 lignes de code apparaissent : une pour paramétrer la méthode, une pour construire le graphe des variables et une pour construire le graphe des individus ici avec des ellipses et la fonction plotellipses.

Enfin on peut quitter l'application en cliquant sur ce bouton « quitter l'application ». Mon objet res contient les lignes de code pour paramétrer la méthode et construire le graphe des individus et celui des variables. Je peux aussi retrouver l'application exactement dans l'état dans lequel je l'avais laissée précédemment. Il suffit que je fasse Factoshiny (res). Vous voyez qu'on retrouve tout le paramétrage. Je peux donc à nouveau modifier mes graphes. Et je peux quitter à nouveau l'application et fermer.

Vous avez vu les principales fonctionnalités de la fonction Factoshiny, n'hésitez pas à tester les fonctionnalités disponibles. A vous maintenant de mettre en œuvre des ACP avec FactoMineR et Factoshiny.