Transcription de l'audio de la vidéo sur l'Analyse Factorielle Multiple

L'objectif de cette vidéo est de montrer comment faire une analyse factorielle multiple, une AFM, avec FactoMineR et comment améliorer les graphiques construits par défaut.

Nous allons utiliser le jeu de données suivant où 21 vins ont été décrits par un panel d'experts à l'aide de 29 descripteurs sensoriels. Deux variables qualitatives, l'origine des vins et le sol, ont été également enregistrées. Nous avons donc en ligne 21 vins et en colonnes 31 variables, 2 qualitatives et 29 quantitatives qui peuvent être regroupées en 5 groupes de variables : un groupe d'odeur, un groupe de visuel, un groupe sur les variables de goût, un groupe sur l'odeur après agitation, et un groupe sur l'appréciation olfactive et gustative. Les groupes d'odeur, visuel, goût et odeur après agitation ont été considérés comme des groupes actifs, les groupes d'origine et d'appréciation comme des groupes supplémentaires.

Ouvrons R ou Rstudio.

Nous allons en fait utiliser le jeu de données wine qui est le jeu de données exemple de la fonction MFA. Nous chargeons le package Factoshiny puis le jeu de données en faisant data(wine). Ensuite, nous lançons la fonction Factoshiny sur le jeu de données wine grâce à Factoshiny (wine).
La fenêtre s'ouvre dans le navigateur par défaut. On choisit donc ici la méthode d'analyse factorielle multiple. La première chose à faire est de définir chaque groupe de variables. Il y a 2 possibilités pour définir les groupes. Une première façon de faire consiste à construire les groupes 1 à 1, variable par variable. Ceci peut être intéressant quand il y a peu de variables et quand les variables du jeu de données ne sont pas rangées dans l'ordre des groupes. Cependant, quand il y a beaucoup de variables ou beaucoup de groupes, cette façon de faire est fastidieuse. On préférera alors définir les groupes en donnant le nombre de variables appartenant à chacun des groupes, et en précisant la nature de chacun des groupes.

Voyons comment définir les groupes à partir de l'interface sur un exemple simple. Je vais construire un premier groupe de variables avec les variables qualitatives, je sélectionne donc qualitatif puis les variables, et je peux donner un nom à ce groupe. Je peux préciser que ce groupe est illustratif et je peux nommer le groupe. Je peux ensuite construire un groupe de variables quantitatifs avec 3 variables quantitatifs, puis un 3ème groupe de variables quantitatives qui contiendra 4 variables. Si je ne veux plus ajouter de groupes, je dois alors valider les groupes.

Revenons à notre exemple complet qui contient plus de variables. Nous allons maintenant utiliser la seconde stratégie pour définir les groupes. Il nous faut préciser dans un premier temps le nombre de variables par groupe, ces nombres étant séparés par des espaces ou des virgules. Dans l'exemple wine, le premier groupe est constitué des 2 premières variables, le 2ème groupe contient les 5 suivantes, puis le 3ème les 3 suivantes, le 4ème les 10 suivantes, le 5ème groupe les 9 suivantes et le dernier groupe contient les 2 dernières

variables. On précise ensuite la nature des variables de chaque groupe avec "c" pour continues ou quantitatives, "s" pour continues ou quantitatives mais les variables sont réduites ("scaled" en anglais); "n" pour les groupes de variables nominales (ou qualitatives) et "f" pour les tableaux de fréquences. Donc, dans l'exemple, le premier groupe est composé de variables qualitatives, tandis que tous les groupes suivants contiennent des variables quantitatives avec un "s" ici qui signifie qu'on va réduire les variables dans chaque groupe.
On peut préciser le nom des groupes de variables, séparés par des espaces. Si on ne met rien, on aura les noms group1, group2, etc. Ensuite on précise les groupes de variables supplémentaires s'il y en a : dans l'exemple, les groupes 1 et 6, c'est-à-dire le groupe d'origine et le groupe des variables d'ensemble sont des groupes supplémentaires qui ne participent pas à la construction des axes. On valide alors la construction des groupes.

Je peux ensuite aller dans paramètres et préciser si certains individus sont supplémentaires, ou encore préciser comment gérer les données manquantes si des données manquantes sont présentes dans le jeu de données. On pourra imputer par la moyenne de la variable pour les variables quantitatives ou bien par la proportion pour les variables qualitatives, ou encore en utilisant un modèle d'AFM à 2 dimensions.
On peut ensuite travailler les graphes et analyser les sorties.

On peut commencer par ouvrir l'onglet Valeurs qui contient les résultats numériques. On trouve un tableau avec les valeurs propres et les pourcentages d'inertie associés à chaque dimension. La première dimension récupère 49% de l'information, *i.e.* 49 % de l'inertie, et la deuxième dimension 19%. Ensuite on a des résultats sur les groupes de variables donc les groupes actifs d'abord avec les coordonnées des groupes, les contributions de chacun des groupes à la construction de la première dimension et les qualités de représentation sur la première dimension. Puis, les mêmes résultats sur la deuxième dimension (coordonnées, contributions et cosinus carrés) puis la 3ème dimension. Ensuite, on a les résultats sur les groupes supplémentaires avec les coordonnées et les cosinus carrés. On n'a pas de contribution puisque ce sont des groupes qui ne contribuent pas à la construction des axes. Le tableau suivant fournit les résultats sur les individus, par défaut les dix premiers individus. Si on veut les résultats sur tous les individus, on modifie la valeur en mettant une valeur élevée. Cette valeur vaut pour tous les tableaux : des individus mais aussi des variables, etc. On retrouve les coordonnées, les contributions et les cosinus carrés, d'abord sur la dimension 1, puis 2, puis 3. Ensuite, nous avons les résultats sur les variables quantitatives actives, donc les résultats pour les dix premières variables avec là encore coordonnées, contributions, cos². Pour les variables quantitatives supplémentaires, on a juste les coordonnées et les cos²; là encore ce sont des variables qui n'ont pas contribué à la construction des axes. Il n'y a pas de variables qualitatives ici qui sont actives donc on a des résultats uniquement pour les variables qualitatives supplémentaires et plus particulièrement pour les modalités des variables qualitatives supplémentaires. Donc on a la coordonnée, le cosinus carré, pas de contribution bien entendu et une valeur-test notée `v.test`. La valeur-test permet de savoir si la coordonnée de la modalité est significativement différente de 0 ou non. Plus précisément, la valeur-test correspond à la transformation d'une probabilité en quantile de loi Normale. Si la probabilité est inférieure à 5% alors la valeur absolue de la valeur-test sera supérieure à 1.96. Le signe de la valeur-test indique si la coordonnée de la modalité est inférieure (pour un signe négatif) ou supérieure (pour un signe positif) à 0. Par exemple, la modalité `Env4` a une coordonnée significativement différente de 0 sur la

deuxième dimension; et supérieure à 0 sur cette 2ème dimension. Voilà pour les principaux résultats numériques obtenus à l'issue d'une AFM.
Voyons maintenant les graphes. Plusieurs graphes sont dessinés. Le nombre de graphes ainsi que les graphes construits dépendent de la présence ou non de groupes de variables quantitatives et de groupes de variables qualitatives. Décrivons maintenant 1 à 1 les graphes obtenus dans notre exemple.

On a le graphe avec les individus et les modalités des variables qualitatives actives et supplémentaires. Les points individus correspondent aux individus moyens, c'est-à-dire aux individus vus par l'ensemble des groupes de variables actifs. On dira par exemple que les vins T1 et T2 ont globalement des profils sensoriels très proches quand on prend en compte tous les points de vue sensoriels (visuel, olfactif, gustatif, etc.). Les modalités sont au barycentre des points qui la prenne. On peut rendre les modalités invisibles par exemple. On peut aussi demander à avoir les points partiels. Les points partiels sont habillés par les couleurs utilisées dans le graphe des groupes. Pour le vin 1VAU, le point rouge représente comment est vu le vin 1VAU par les variables d'olfaction uniquement et le point vert comment ce vin est vu par rapport aux variables visuelles uniquement.

Il est possible de ne pas dessiner tous les points et de sélectionner certains individus en fonction de leur contribution ou de leur qualité de représentation. On peut par exemple sélectionner uniquement les vins bien projetés sur le plan, qui ont un cosinus carré supérieur à 0.4. Je peux également habiller les individus en fonction d'une variable qualitative, par exemple de la variable Soil. Si je veux récupérer les lignes de code correspondant à un graphe, je vais cliquer sur le bouton ligne de code de l'ACM et récupérer la ligne de code.

On peut aussi construire un graphe avec les modalités et leurs points partiels. Ici il n'y a pas de modalités actives donc seules les modalités supplémentaires sont représentées. Pour chaque modalité, le point moyen ainsi que ses points partiels sont représentés. Points moyens et partiels s'interprètent ici comme pour le graphe des individus. On pourra dire par exemple que les vins sur le sol référence et ceux sur le sol Env4 sont visuellement perçus de la même façon mais ils sont très différents sur les descriptions olfactives et gustative.

Comme dans notre analyse certains groupes sont constitués de variables quantitatives, on trouve un graphe des variables avec le cercle des corrélations. Dans ce graphe, les variables sont coloriées en fonction de leur groupe d'appartenance. Donc une même couleur pour un même groupe. Ce graphe des variables s'interprète simultanément avec le graphe des individus comme pour une ACP. Donc par exemple, les vins sur la droite du graphe des individus sont des vins intenses au niveau de l'arôme et du goût. Et les vins sur le haut du graphe sont épicés (spice en anglais). On peut retravailler ce graphe et sélectionner les variables en fonction de leur contribution par exemple. Ici on prend les 5 variables qui ont le plus contribué à la construction des axes. Cela permet d'avoir des graphes avec des libellés qui vont moins se chevaucher. Donc des graphes qui sont plus lisibles si on a beaucoup de variables. Souvent il est intéressant de faire une sélection car les variables proches du centre du cercle, *i.e.* avec des flèches très courtes, sont des variables peu intéressantes à interpréter car mal projetées. On s'intéressera souvent aux variables les mieux projetées, qui ont donc une coordonnée élevée, ou qui ont fortement contribué (c'est la même information) à la construction des axes.

On trouve ensuite le graphe des groupes de variables avec des triangles pleins pour les groupes actifs et des triangles vides pour les groupes illustratifs. Les couleurs des groupes sont les mêmes que dans les autres graphes. Si un groupe a une coordonnée élevée sur une dimension, alors cette dimension est également présente dans le groupe de variables. Autrement dit, ce groupe sépare les individus comme peut le faire l'AFM avec cette dimension. Si deux groupes ont des coordonnées élevées et proches sur plusieurs dimensions, alors ils induisent une structure proche sur les individus.

On trouve enfin le graphe avec les axes partiels. Pour chaque groupe de variables on a fait une analyse : pour les variables quantitatives, une ACP, pour les variables qualitatives une ACM, et on a projeté en supplémentaire les dimensions de ces ACP et ACM sur les dimensions de l'AFM. Donc par exemple, pour le groupe de visuel, la première dimension du visuel est très liée à la première dimension de l'AFM alors que la deuxième dimension est un peu moins liée à la 2ème dimension de l'AFM. Pour le groupe de variables qualitatives origine, on a projeté les dimensions de l'ACM. Ici, les 2 premières dimensions de chaque groupe sont représentées, mais on peut choisir d'en représenter 3.

Par défaut, le plan 1-2 est fourni mais on peut construire le plan 3-4.

Voilà pour les principales sorties graphiques.

Il y a aussi un onglet qui fournit une description automatique des dimensions factorielles exactement comme pour les résultats d'une ACP ou d'une ACM. La première dimension factorielle est donc décrite par les variables quantitatives actives et supplémentaires les plus liées (i.e. corrélées positivement et négativement) à la première dimension de l'AFM puis par les variables qualitatives et les modalités les plus liées. Par défaut les 3 premières dimensions factorielles sont décrites.

Il est possible, à l'issue de l'AFM, de réaliser une classification. Cette classification utilisera donc la pondération des groupes dans le calcul des distances entre individus. Il faut alors choisir le nombre de dimensions de l'AFM qui seront conservées pour faire l'a classification. Nous ne montrons pas ici la classification car une autre vidéo présente en détail cette méthode.

Comme dit précédemment, le bouton « lignes de codes de l'AFM » récupère les lignes de codes de l'AFM pour mettre en œuvre la méthode et construire les graphes à l'identique. En cliquant sur lignes de codes de l'AFM, les lignes de code apparaissent : une pour paramétrer la méthode, d'autres pour construire les graphes.

Vous avez vu les principaux graphes et les principaux indicateurs de l'AFM. A vous de mettre en œuvre cette méthode grâce à Factoshiny.