Transcription de l'audio de la vidéo sur l'ACM avec FactoMineR

Nous allons voir comment mettre en œuvre l'ACM sous FactoMineR. Pour ce faire, nous allons analyser un questionnaire sur la consommation de thé. Ce questionnaire a été rempli par 300 personnes et contient trois types de variables. 18 variables concernent le mode de consommation: par exemple prenez-vous du thé au petit-déjeuner? oui/non? Au goûter ? oui/non? Sous quelle forme? En sachet, en vrac ou les 2? Ca concerne les 18 premières variables du jeu de données. Ensuite il y avait des questions sur l'image que les enquêtés ont du thé et des questions sur le signalétique. Ce sont les questions qui sont en vert ici. Sur le signalétique, par exemple le sexe, la CSP est-ce que les personnes sont sportives ou non? l'âge, en tant que variable qualitative, codée Age_q, l'âge est ici codé en classes d'âge, et puis des questions sur l'image du thé: est-ce que le thé est pour vous bon pour la santé? Est-ce que c'est diurétique? etc. Et puis un dernier type de variable qui est une variable quantitative: l'âge.

Dans un premier temps nous allons importer le jeu de données directement depuis internet. Le jeu de données est disponible sur le site Internet de FactoMineR et c'est le fichier thé.csv. Je précise que le nom des variables est disponible avec header=TRUE, que le séparateur de colonnes est le ";".

Voyons maintenant comment lancer l'ACM sur le jeu de données thé grâce à Factoshiny. On écrit Factoshiny(the) ce qui ouvre l'interface graphique dans le navigateur. 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 clique sur la méthode choisie à savoir l'analyse des correspondances multiples. Une nouvelle fenêtre s'ouvre.

Commençons par paramétrer la méthode. La variable quantitative "age" est considérée comme supplémentaire puisqu'elle est quantitative ; si je la supprime ici elle disparaîtra de l'analyse. Je vais préciser que les variables concernant l'image du thé et le signalétique seront supplémentaires (je dois les sélectionner une à une). Ici, il n'y a pas de données manquantes dans le jeu de données. S'il y en avait, on pourrait gérer les données manquantes selon plusieurs options : en ajoutant une modalité NA pour chaque variable contenant des valeurs manquantes (c'est l'option par défait) ; en imputant les valeurs manquantes du tableau disjonctif complet par la proportion de la modalité sur les valeurs observées. Cette méthode est très rapide mais pas recommandée. Il est possible d'imputer le tableau disjonctif par un modèle d'ACM à 2 dimensions, ce qui est plutôt un bon compromis dans la plupart des situations. Et enfin on peut imputer par un modèle d'ACM à k dimensions. Le nombre k est dans un premier temps estimé par validation croisée, ce qui peut être long sur de gros jeux de données. Mais le nombre de dimensions est alors optimal pour le modèle d'ACM qui servira à imputer le tableau disjonctif. Une fois le tableau disjonctif imputé, l'ACM est alors construite à partir de ce nouveau tableau disjonctif.

Une fois le paramétrage terminé, je soumets ou alors je sors de cet onglet paramètres. Cela soumet la méthode avec le nouveau paramétrage.

Nous avons différents graphes : un graphe avec les individus et les modalités, actives en rouge, et supplémentaires en vert (ce graphe est très chargé, nous verrons comment l'améliorer). Un graphe avec le carré des liaisons. Pour les variables qualitatives actives et supplémentaires, on utilise le eta² entre la variable qualitative et l'axe; pour la variable age on utilise le R^2 de la régression de la variable age en fonction de chacune des dimensions. Et enfin on obtient un graphe des variables quantitatives supplémentaires (ici, il n'y en a qu'une seule), avec le cercle des corrélations.

Avant de voir comment améliorer les graphes, voyons les résultats de l'ACM dans le second onglet. Nous avons d'abord un rappel de la ligne de code utilisée, ensuite un tableau avec les valeurs propres et les pourcentages d'inertie associée à chaque axe, ensuite les résultats sur les individus. Par défaut les résultats pour les 10 premiers éléments sont données. Si on met une valeur plus grande on verra plus de résultats. Pour les individus, nous avons leurs coordonnées sur la première dimension, la contribution à la construction de l'axe, en pourcentage, par exemple le 1er individu a contribué à 0.6% à la construction de l'axe; ce qui est assez peu mais il y a 300 individus. Et puis la qualité de représentation mesurée par le cosinus carré. Nous avons les résultats sur les dimensions 1, 2 et 3. Nous avons ensuite les résultats sur les modalités avec les coordonnées des modalités, les contributions, la qualité de représentation par le cosinus carré et une valeur-test. La valeur-test ici suit une loi Normale et donc les valeurs-tests qui sont inférieures à -2 ou supérieures à 2 seront considérées comme significatives, c'est-à-dire que la modalité a une coordonnée significativement différente de 0. Ce qui est utile pour savoir quelles sont les modalités qui ont des valeurs extrêmes sur les dimensions.

Par exemple, ici, petit-déjeuner a une valeur extrême et significativement négative sur la 1ère dimension. Nous avons les résultats sur la 1ère dimension, la 2ème dimension et la 3ème dimension. Nous avons les résultats sur les variables globalement, à savoir le rapport de corrélation de la variable sur chacune des dimensions 1, 2 et 3. C'est juste le rapport de corrélation entre la dimension et la variable. Ce sont ces coordonnées qui permettent de construire le graphe du carré des liaisons. Ensuite, si nous avons des modalités supplémentaires, nous avons les résultats sur les modalités supplémentaires avec la coordonnée, la qualité de représentation, pas de contribution puisque les modalités supplémentaires ne contribuent pas à la construction des axes, et la v-test. Enfin nous avons même chose, un tableau avec les rapports de corrélation des variables qualitatives supplémentaires. Et puis enfin nous avons des résultats sur les variables quantitatives supplémentaires avec la coordonnée des variables sur chaque axe, c'est-à-dire le coefficient de corrélation entre l'axe et la variable quantitative.

Nous pouvons décrire les dimensions, ce qui est très utile quand nous avons beaucoup de variables. Nous voyons par exemple que la 1ère dimension n'est pas caractérisée par des variables quantitatives; il n'y a aucune variable quantitative qui caractérise la 1ère dimension. Par contre la 1ère dimension est liée, de façon significative, à différentes variables qualitatives. La variable la plus liée est le lieu d'achat, ensuite salon de thé, etc.

Les variables sont triées des plus liées au moins liées et seules les variables qui ont un rapport de corrélation significativement différent de 0 sont conservées dans les résultats. Nous avons ensuite les tableaux des modalités avec chaque modalité et sa coordonnée sur l'axe et la probabilité critique du test "est-ce que cette coordonnée est significativement différente de 0 ou non?". Nous pouvons noter que nous avons des variables actives et supplémentaires dans la description des axes, de même des

modalités actives et supplémentaires. Nous pouvons voir les résultats sur la 2ème dimension : la variable âge est significativement corrélée à cette dimension.

Revenons sur les options graphiques pour voir comment améliorer la lisibilité des graphiques. Nous pouvons construire un graphe avec seulement les individus. Je précise alors que je ne dessine que les individus. Je peux aussi supprimer les libellés de ces individus. Au contraire, je peux ne conserver que les modalités actives. 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 choisir de ne représenter que les individus qui ont une qualité de représentation élevée ou bien dessiner ceux qui ont le plus contribué à la construction du plan. On peut faire de même pour les modalités. On voit ici un graphe avec les 10 modalités qui ont le plus contribué à la construction du plan.

Je peux aussi colorier les individus en fonction de leur qualité de représentation ou de leur contribution. Par exemple, en fonction de leur contribution, on voit que les individus les plus extrêmes contribuent le plus. On peut aussi dessiner les individus en fonction d'une variable. Par exemple, en fonction de la variable where. On a une couleur différente pour chacune des 3 modalités de cette variable. On retrouve que cette variable est liée aux 2 premières dimensions puisque les sous-nuages sont bien séparés. On peut aussi construire des ellipses de confiance autour de chacune des 3 modalités. On voit que les ellipses de confiance sont toutes petites. En effet le nombre d'individus est important ici et les sous-populations sont assez séparées. Les ellipses ne se chevauchent pas, ce qui indique que les sous-populations sont significativement séparées.

Enfin, nous pouvons colorier toutes les modalités d'une même variable avec des couleurs différentes. Cela permet de mieux voir les modalités d'une même variable. Ce graphe est très utile quand il y a peu de variables et que les variables ont un nombre de modalités important. Les couleurs utilisées sur le graphe des modalités sont simultanément utilisées sur le graphe des variables. Ce graphe des variables donne les carrés des liaisons et indique quelles variables sont globalement liées aux différentes dimensions. On peut également améliorer ce graphe en cliquant sur Variables. Là encore, on peut rendre invisible les variables supplémentaires par exemple.

Bien entendu, nous pouvons voir les axes 3 et 4 et pas seulement les axes 1-2.

A l'issue de l'ACM, 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'ACM, 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'ACM, ce qui revient à faire une classification sur les données initiales, l'ACM ayant servi de prétraitement pour passer de données qualitatives à des données quantitatives. 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'ACM, i.e. une interprétation automatique simple des résultats de l'ACM en anglais ou en français. Ce rapport automatique peut utiliser des

graphes suggérés par l'analyse ou les graphes que l'on vient de travailler. Dans un premier temps, il est intéressant de voir les graphes suggérés par la méthode. On peut récupérer ce rapport automatique au format Rmarkdown, au format html ou au format word.

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

Enfin on peut quitter l'application en cliquant sur ce bouton « quitter l'application ». L'objet res donne les lignes de code pour paramétrer la méthode et construire les différents graphes. On peut également retrouver l'application exactement dans l'état dans lequel on l'avait laissée précédemment en faisant Factoshiny(res). On peut donc à nouveau modifier les graphes… et quitter à nouveau l'application.

A vous maintenant de mettre en œuvre des ACM avec FactoMineR et Factoshiny.