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

Nous allons voir comment réaliser une analyse des correspondances avec FactoMineR. Pour ce faire, nous allons travailler sur les données du nombre de naissances en 2003 en fonction de l'âge du père et de l'âge de la mère. En lignes nous avons les différentes tranches d'âge: les mères âgées de moins de 20 ans, entre 20 et 24 ans, 25-29 etc. et en colonnes les pères âgés de moins de 20 ans, 20 à 24, etc. Donc dans la première case par exemple, 2085 correspond au nombre d'enfants nés d'une mère de moins de 20 ans et d'un père de moins de 20 ans. Dans ce tableau certains effectifs sont vraiment très petits. Il s'agit des nombres de naissances pour des mères âgées de 45 à 49 ans et de 50 à 60 ans. D'un point de vue pédagogique, nous avons voulu regrouper ces deux modalités en une modalité mère de + de 45 ans. Nous allons utiliser en actifs les modalités *mère moins de 20 ans*, *de 20 à 24*, jusqu'à *40 à 44 ans* et *mère de plus de 45 ans*. Et nous allons utiliser les lignes *mères de 45 à 49 ans* et *mères de 50 à 60 ans* comme des lignes illustratives. De même les pères âgés *de 55 à 59 ans* et pères *plus de 60 ans* sont considérés comme des colonnes illustratives. Nous allons donc considérer comme actifs les colonnes pères *moins de 20 ans*, jusqu'à père de *50 à 54 ans* et *pères + 55ans*. Les couleurs qui sont ici dans les libellés des lignes et des colonnes diffèrent selon que les lignes et colonnes sont actives ou supplémentaires. Ce sont les couleurs que nous retrouverons ensuite dans les graphes de l'AFC.

Commençons par importer le jeu de données. Le jeu de données est disponible via le lien suivant. Nous allons l'importer avec la fonction read.table, en précisant que le nom des colonnes est disponible avec header=TRUE, que le séparateur de colonnes est le ";", que le nom des lignes est disponible dans la première colonne du jeu de données; check.names=FALSE indique que le nom des colonnes doit être pris tel quel (sinon R modifie les espaces en mettant des "."). Nous pouvons vérifier que les données sont bien importées : ici toutes les variables sont bien quantitatives.

Voyons maintenant comment lancer l'AFC 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'AFC, il suffit de lancer 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 naissances. L'interface graphique s'ouvre dans le navigateur par défaut. 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 lignes et

colonnes actives et supplémentaires, mais également la gestion des données manquantes si des données manquantes sont présentes dans le jeu de données.

Je vais maintenant préciser que les lignes *mères de 45 à 49 ans* et *mères de 50 à 60 ans* sont illustratives et les colonnes *pères âgés de 55 à 59 ans* et *pères plus de 60 ans* sont illustratives. Dans notre jeu de données, il n'y a ni variables quantitatives supplémentaires, ni variables qualitatives supplémentaires. Qu'est-ce que serait des variables quantitatives ou qualitatives supplémentaires. Prenons un autre jeu de données qui permet d'expliciter plus facilement ces variables. On a le tableau de contingence croisant des auteurs en lignes et les mots qu'ils utilisent dans leur texte en colonnes. On peut ajouter comme variable quantitative l'année de naissance des auteurs, ce qui permettrait par exemple de mettre en évidence une évolution temporelle. L'année est bien ici une variable quantitative, bien différente d'une donnée de comptage. Il ne s'agit donc pas d'une colonne supplémentaire. Pour une variable qualitative supplémentaire, on peut imaginer une variable qui organise les auteurs par courant littéraire. La variable courant littéraire est qualitative et prend plusieurs modalités. Notez que ces variables supplémentaires concernent uniquement les lignes du tableau. Pour les variables quantitatives, on les représente sur un graphe avec le cercle des corrélations, en calculant le coefficient de corrélation entre la variable quantitative et les coordonnées des lignes sur les axes. Pour les variables qualitatives, on positionne les modalités d'une variable qualitative au barycentre des lignes qui prennent cette modalité. A chaque fois, le poids de la ligne est pris en compte. Dans notre exemple, il n'est pas proposé de choisir de variables qualitatives supplémentaires car il n'y a aucune variable qualitative dans le jeu de données. S'il y en avait, on aurait une rubrique pour choisir des variables qualitatives.

Si nous avions des données manquantes, nous aurions plusieurs options pour les gérer : la première option consiste à mettre en supplémentaire les lignes et colonnes qui ont au moins une donnée manquante (c'est ce qui est fait par défaut). La deuxième option consiste à imputer le jeu de données en utilisant le modèle d'indépendance : pour une cellule manquante, on calcule le produit de la somme de sa ligne par la somme de sa colonne, divisé par la somme du tableau. On met à jour toutes les cellules manquantes puis on itère jusqu'à convergence car les sommes en lignes et colonnes bougent. Enfin, il est possible d'imputer le tableau par un modèle d'AFC à 2 dimensions. Cette stratégie est certainement la meilleure dans de nombreuses situations. Ensuite, une fois les données manquantes imputées, l'AFC est construite sur le tableau complété.

Revenons à notre jeu de données sans données manquantes.

L'AFC est réalisée et le graphique de la représentation simultanée est fourni. Mais voyons dans un premier temps les résumés des principaux résultats. Nous avons, dans cet onglet, un listage avec les principaux résultats de l'analyse. La première ligne nous rappelle la commande qui a été lancée pour réaliser l'AFC. Ensuite les résultats du test du Chi² sur les variables avec uniquement les lignes et les colonnes actives montrent que la statistique du Chi² est très grande et que la probabilité critique est très inférieure à 5% et indique qu'il y a une liaison significative entre les 2 variables *âge de la mère* et *âge du père*. Nous avons ensuite un tableau avec les valeurs propres et les pourcentages d'inertie associés à chaque dimension. Puis les résultats sur les lignes actives, avec la coordonnée des lignes sur la première dimension, la contribution de la ligne à la construction de la première dimension et la qualité de représentation sur la première dimension. Ensuite nous avons les résultats sur la deuxième dimension puis sur la troisième dimension.

Même chose pour les colonnes actives. Nous avons les cordonnées, les contributions et les cosinus carrés sur la première, deuxième puis troisième dimension. Nous avons ensuite les résultats sur les éléments supplémentaires. Les lignes d'abord supplémentaires avec la coordonnée et la qualité de représentation. Nous n'avons pas de contribution puisque ces lignes ne contribuent pas à la construction des axes. Et même chose pour les colonnes: les coordonnées et qualité de représentation sur les dimensions de 1 à 3.

Nous pouvons revenir sur les graphiques et améliorer le graphique en diminuant la taille de la police et en modifiant par exemple le titre. Avec une taille de police plus petite, les libellés se superposent beaucoup moins et le graphique est beaucoup plus lisible. Nous pouvons aussi rendre invisible les lignes supplémentaires et colonnes supplémentaires et ne conserver que les lignes et colonnes actives. Nous pouvons également ne mettre des libellés que pour les éléments qui sont bien représentés. Donc ici les lignes qui ont une qualité de représentation suffisante avec un cosinus carré supérieur à 0.7 sur le plan, et les colonnes avec un cosinus carré supérieur à 0.7 sur le plan. Les lignes et colonnes qui ont des libellés ici sont bien projetées sur ce plan, les autres n'ont pas de libellé et sont écrits avec transparence. Ceci est très utile quand nous avons énormément d'éléments pour faire le tri dans les éléments à commenter dans un graphe. Nous pouvons aussi dessiner les libellés en fonction des contributions. Par exemple, les 4 lignes et les 3 colonnes qui ont le plus contribué à la construction du plan. Il est également possible de dessiner tous les éléments et de les colorier en fonction de leur qualité de représentation sur le plan, ou de leur contribution à la construction du plan. J'enlève donc la sélection par contribution et choisis de colorier les libellés en fonction de la contribution par exemple. On remarquera que, contrairement à l'ACP, les éléments (lignes ou colonnes) qui contribuent le plus ne sont pas nécessairement ceux qui sont les plus éloignés du barycentre. Le poids de l'élément entre en jeu ici et une ligne avec un fort effectif peut fortement contribuer, même si elle n'est pas très éloignée du barycentre.

Nous pouvons également tracer des ellipses de confiance autour de la position des lignes et des colonnes. Le principe de la construction des ellipses est le suivant : le tableau avec les éléments actifs est pris comme référence. Ensuite, de nouveaux tableaux de données sont construits en tirant N valeurs dans une distribution multinomiale avec des fréquences théoriques égales aux valeurs dans les cellules du tableau divisées par N. L'idée est alors de fournir une zone de confiance à 95% sur la position du point. Ici, les ellipses sont vraiment toutes petites car les effectifs dans le tableau de données sont très grands, plusieurs dizaines de milliers, et donc la position de chaque point est très stable. Quand les effectifs sont plus petits, les ellipses sont beaucoup plus larges et renseignent sur la possibilité d'interpréter ou non une différence de position entre 2 lignes ou entre 2 colonnes.

Enfin, nous pouvons bien sûr construire des graphes avec d'autres dimensions, par exemple dessiner le plan 3-4.

A l'issue de l'analyse des correspondances, nous pouvons construire une classification. Après une analyse des correspondances, il est possible de construire une classification sur les lignes ou sur les colonnes. Ce choix pourra être fait dans l'application de la classification. Nous devons préciser ici le nombre de dimensions de l'AFC qui seront conservées pour construire la classification.

Il est aussi possible d'obtenir un rapport sur les résultats de l'AFC, i.e. une interprétation automatique simple des résultats de l'AFC 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.

Enfin il y a un bouton « lignes de codes » de l'AFC qui récupère les lignes de codes de l'AFC pour mettre en œuvre la méthode et reconstruire les graphes à l'identique. Donc si je clique sur lignes de codes de l'AFC, les 2 lignes de code apparaissent ici : une pour paramétrer la méthode et une pour construire le graphe.

Enfin on peut quitter l'application en cliquant sur ce bouton « quitter l'application ». Si j'affiche l'objet res, je retrouve les lignes de code qui permettent de paramétrer la méthode et de construire le graphe. Je peux également retrouver l'application dans l'état dans laquelle je l'avais laissée en tapant Factoshiny(res). Vous voyez qu'on retrouve l'application exactement dans l'état où elle était précédemment. Je peux donc à nouveau modifier mes graphes. Et je peux quitter à nouveau l'application et fermer.

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