Transcription de l'audio de la vidéo de conclusion présentant la démarche en analyse des données

Nous allons voir dans cette dernière vidéo un récapitulatif des méthodes que nous avons vues durant ce MOOC. Pour ce faire, nous allons détailler les différentes questions que l'on se pose lorsqu'on est amené à analyser un jeu de données.

La première question à se poser est : y a-t-il des groupes de variables ? Donc est-ce que le jeu de données est structuré avec différentes sources d'information, chaque source d'information contenant plusieurs variables. Si oui, on utilise des analyses qui prennent en compte ces groupes et donc on peut construire une Analyse Factorielle Multiple.

Deuxième question à laquelle on doit répondre : quel est le type d'information auquel on est confronté ? Est-ce qu'on a un ou plusieurs tableaux de contingence, donc des tableaux avec des lignes et des colonnes et dans une cellule du tableau un effectif. Dans ce cas on s'oriente vers une analyse des correspondances (AFC) si on a un tableau ou vers une analyse factorielle multiple sur tableaux de contingence si on veut comparer plusieurs tableaux de contingence. Autre type de tableaux : des tableaux avec des individus statistiques en lignes et des variables en colonnes. Les individus sont décrits par plusieurs variables. Et donc là on a différentes méthodes selon la nature des données : l'ACP si les variables sont quantitatives, l'ACM si les variables sont qualitatives, l'AFDM si les données sont mixtes ou encore l'AFM si les variables sont structurées en groupes. Décrivons rapidement et succinctement une méthode d'analyse factorielle que nous n'avons pas vue dans ce MOOC et qui permet de prendre en compte simultanément des variables quantitatives et qualitatives. Il s'agit de l'analyse factorielle de données mixtes (AFDM) qui a pour principe d'équilibrer l'influence de toutes les variables dans l'analyse, que les variables soient quantitatives ou qualitatives. Les graphes et résultats numériques sont un mixte des résultats d'ACP et d'ACM : on dessinera le graphe des individus et modalités construit en ACM et le graphe du cercle des corrélations de l'ACP pour les variables quantitatives. Les différentes fonctions de FactoMineR, écrites en orange, sont CA pour l'analyse des correspondances, MFA pour l'analyse factorielle multiple sur tableaux de contingence, PCA pour l'ACP, MCA pour l'analyse des correspondances multiples, FAMD pour l'analyse factorielle de données mixtes et MFA pour l'analyse factorielle multiple.

Troisième question à laquelle il faut répondre : quels sont les éléments actifs ? A travers éléments actifs on entend: quels sont les éléments qui doivent participer à la construction des axes factoriels ? Quels sont les éléments à partir desquels nous calculons les distances entre individus ou entre lignes? Nous allons avoir un tableau avec ici en bleu les éléments actifs, en vert les colonnes illustratives et en rose les lignes illustratives. Donc selon les méthodes, nous utilisons les arguments suivants: ind.sup pour considérer que ce sont des individus supplémentaires, quanti.sup pour des variables quantitatives supplémentaires, quali.sup pour des variables qualitatives supplémentaires et row.sup, col.sup pour des lignes ou des colonnes supplémentaires en AFC et enfin group.sup pour des groupes de variables supplémentaires en AFM.

Une fois que nous avons défini les éléments actifs, il est nécessaire de savoir quelle est la nature des variables actives. La nature peut-être de deux types: quantitative ou qualitative. Si les variables sont quantitatives alors nous avons affaire à une ACP. Si les variables sont qualitatives, deux choses: si nous avons uniquement deux variables, dans ce cas il est conseillé de construire le tableau croisé de ces deux variables avec en lignes les modalités d'une variable, en colonnes les modalités de l'autre et nous construisons le tableau de contingence et analysons ce tableau par une AFC. Et si nous avons plus de deux variables qualitatives nous considérons l'ACM. Si les variables actives sont à la fois quantitatives et qualitatives, nous pouvons utiliser l'analyse factorielle de données mixtes. En présence de groupes de variables, on met en œuvre une analyse factorielle multiple, en précisant le type de chaque groupe de variables.

Autre question importante qui concerne les variables quantitatives quand on réalise une ACP ou une AFM : doit-on réduire ou non les variables? Nous avons vu que si les variables étaient dans des unités différentes, il est nécessaire de réduire. Maintenant si toutes les variables actives sont dans la même unité, on peut réduire ou non les variables. Réduire conduit à accorder la même importance à chaque variable et ne pas réduire conduit à accorder plus d'importance aux variables qui ont une plus forte variance. Donc si c'est une ACP réduite, on utilise scale.unit=TRUE dans la fonction PCA de FactoMineR. Cette question se pose également pour chaque groupe quantitatif de l'AFM. Si on souhaite réduire un groupe de variables on utilise type "s" (pour scaled) et si on ne souhaite pas réduire on utilise le type "c" (pour continuous).

Question importante également : y a-t-il des données manquantes? Si nous avons des données manquantes, il est utile d'utiliser le package missMDA. Ce package permet de compléter le jeu de données. Ensuite on peut réutiliser les fonctions classiques d'ACP, d'ACM, d'AFDM ou d'AFM de FactoMineR, sur le tableau complété, le tableau imputé.

Donc voici les 6 questions importantes auxquelles il faut répondre avant de réaliser l'analyse. Ensuite, on peut lancer l'analyse factorielle. Selon le type de données, soit une ACP, soit une analyse des correspondances, soit une analyse des correspondances multiples, soit une analyse factorielle de données mixtes, soit une analyse factorielle multiple.

Une fois l'analyse lancée, on construit les graphes et complète l'interprétation par les résultats numériques. Pour les résultats numériques, on utilise les différentes fonctions summary : summary.PCA, summary.CA, summary.MCA, summary.MFA. Et puis, si on veut améliorer les graphes qui sont fournis par défaut, on peut utiliser les fonctions plot.PCA, plot.CA, plot.MCA, plot.MFA.

Il est toujours intéressant de décrire les axes factoriels par les variables initiales. Il suffit d'utiliser la fonction dimdesc sur les résultats de l'analyse factorielle.

Eventuellement, à l'issue de l'analyse factorielle, on peut réaliser une classification des individus ou bien une classification des lignes ou des colonnes après une AFC. La classification peut être construite avec la fonction HCPC de FactoMineR, fonction qui permet également de décrire les classes construites à l'aide des variables initiales.

Il est temps maintenant de vous remercier d'avoir suivi ce cours jusqu'au bout.

Merci de votre attention et à bientôt.