Transcription de l'audio du cours d'Analyse Factorielle Multiple

Première partie. Données - problématique

(Diapositives 1 à 6)

Cette semaine nous allons nous intéresser à une méthode permettant d'étudier des tableaux de données plus complexes où un même ensemble d'individus est décrit par des variables structurées en groupes et provenant éventuellement de différentes sources d'information. L'intérêt de la méthode sera d'analyser globalement le tableau de données mais aussi de comparer l'information apportée par les différentes sources d'information.

Diapositive 1 :

Nous vous proposons 4 vidéos de cours qui présentent les principales caractéristiques de l'AFM, l'analyse factorielle multiple.

Diapositive 2 (plan) :

Le plan de l'exposé est le suivant : nous commencerons par décrire la structure des données sur lesquelles on travaille en AFM. Nous détaillerons plusieurs exemples de jeux de données ainsi que leur problématique associée.

Ensuite nous verrons que l'AFM peut être vue comme une ACP particulière sur une matrice pondérée qui permet d'équilibrer l'information apportée par les différents groupes de variables. Nous présenterons la pondération et indiquerons les intérêts de cette pondération. L'AFM fournit des résultats sur les individus et les variables comme l'ACP pour les variables quantitatives ou comme l'ACM pour les variables qualitatives.

La spécificité et la richesse de l'AFM est de prendre en compte plusieurs groupes de variables. Nous verrons comment comparer l'information apportée par chacun de ces groupes, quelle est l'information commune à plusieurs groupes, quelle information est spécifique d'un groupe.

Enfin nous détaillerons plusieurs aides à l'interprétation très utiles pour analyser les résultats d'une AFM.

Diapositive 2 bis (plan) :

Commençons dans cette vidéo par définir sur quel type de données nous travaillons et quelles sont les problématiques associées.

Diapositive 3 :

Présentons tout d'abord le jeu de données à partir duquel nous allons présenter la méthode. Le jeu de données concerne une évaluation sensorielle de 10 vins blancs du Val de Loire. Lors de cette évaluation sensorielle, 5 Vouvray et 5 Sauvignons ont été dégustés et comparés à partir de descripteurs sensoriels comme l'acidité, l'amertume, l'odeur d'agrume, etc.

Pour ce faire, les juges sont amenés à évaluer chacun des vins et à mettre une note sur chaque descripteur. La note varie entre 0 et 10, 0 correspondant à une saveur ou odeur très faible ou inexistante et 10 à une saveur ou odeur très forte. Nous voyons ici les juges en plein travail, disons plutôt en pleine évaluation.

Lorsque tous les juges ont évalué tous les vins selon chaque descripteur, on construit un tableau de moyennes avec en lignes chaque vin, en colonne chaque variable (ou descripteur) et dans une cellule du tableau la note moyenne du vin et du descripteur calculée sur tous les juges.

Le tableau présenté ici concerne l'évaluation faite par un jury d'experts (d'œnologues). Nous pouvons noter à droite du tableau une variable qualitative qui caractérise les vins : la variable cépage qui prend les deux modalités Vouvray et Sauvignon.

Ce tableau rectangulaire, croisant en lignes des individus et en colonnes des variables, peut être analysé par une ACP en mettant en supplémentaire la variable qualitative cépage.

Diapositive 4 :

Mais le problème ici est un peu plus complexe car nous avons obtenu une description sensorielle des mêmes vins par plusieurs jurys. Le jury d'experts dont nous avons vu un extrait du jeu de données mais également un jury d'étudiants et un jury de consommateurs. Les variables utilisées pour décrire les vins ne sont pas les mêmes d'un jury à l'autre, le jury expert a évalué 27 descripteurs sensoriels quand les jurys étudiant et consommateur ont évalué 15 descripteurs sensoriels.

La problématique associé à ce jeu de données est multiple : d'une part décrire les 10 vins d'un point de vue sensoriel grâce aux évaluations des trois jurys, mais également de voir si les jurys décrivent les vins de façon analogue ou bien si certains jurys ont une description spécifique des vins. Pour répondre à cette problématique, on va s'intéresser au tableau de données regroupant les données des trois jurys.

Nous avons de plus à disposition une évaluation de l'appréciation de chacun des vins par un jury de 60 consommateurs. Dans ce tableau, une ligne correspond à un vin et une colonne à un consommateur, et une valeur correspond à la note mise pour un vin par un consommateur. Nous considérons cet ensemble de variables comme illustratif car nous focalisons notre attention sur la description sensorielle des vins. Cependant, nous pourrons relier la description sensorielle et l'appréciation des vins en mettant en supplémentaire ces variables dans l'analyse. De même la variable qualitative correspondant au cépage sera considérée comme qualitative supplémentaire.

Diapositive 5 :

La structure de notre jeu de données exemple est donc la suivante. Un même ensemble d'individus (pour nous les 10 vins) est décrit par plusieurs groupes de variables (dans notre exemple un groupe correspond à un jury). On va considérer que nous avons I individus dans chaque groupe, que nous avons J groupes de variables et que dans chacun de ces groupes nous avons K_j variables. La valeur prise par l'individu i pour la variable k est notée x_{ik}. Dans chaque groupe les variables peuvent être quantitatives ou qualitatives.

On trouve, dans de très nombreux domaines d'application, des tableaux de données de ce type, avec un même ensemble d'individus décrit par plusieurs groupes de variables. En effet, les groupes de variables peuvent correspondre à différentes sources d'information. Voici quelques exemples issus de divers domaines d'application. En génomique, un ensemble de patients atteints de cancer du cerveau est décrit selon 3 sources d'information différentes et éventuellement complémentaires : des mesures d'ADN correspondant à des mesures sur les gènes, des mesures d'expression ou encore des mesures sur les protéines. L'objectif est alors de comprendre quels sont les gènes qui entrent en jeu dans la maladie et qui sont différentiellement exprimés selon que le patient est malade ou non et selon le type de tumeur. L'objectif est également de savoir si les trois sources d'information sont concordantes ou si une des sources apportent des informations différentes.

Un autre exemple concerne une enquête menée auprès de jeunes étudiants. Plusieurs groupes de questions leur ont été posées sur leur consommation de produits tels que l'alcool ou les drogues douces, sur leurs conditions psychologiques, sur la qualité de leur sommeil, et enfin des questions sur leur signalétique. Toutes les questions étaient qualitatives. L'objectif est ici de comprendre globalement le lien entre la qualité du sommeil,

les conditions psychologiques et leur consommation de drogue. On ne s'intéresse pas aux liens variable par variable mais plutôt à des ressemblances plus globales.

Autre exemple où les différents groupes de variables correspondent à différents instants. Cela permet de voir globalement l'évolution de l'ensemble des variables dans le temps. En économie, cela peut correspondre à l'ensemble des indicateurs économiques, d'une période à l'autre ou d'une année à l'autre.

Les exemples de tableaux multiples sont de plus en plus nombreux car d'une part, il est de plus en plus facile de recueillir et stocker des données, et d'autre part, on cherche de plus en plus souvent à analyser des problèmes plus complexes mettant en jeu plusieurs sources d'information.

Diapositive 6 :

Dans la majorité des exemples, la problématique, ou les problématiques, seront les suivantes. Dans un premier temps, étudier et décrire l'ensemble des individus à l'aide de toutes les variables et décrire les relations entre les différentes variables (d'un même groupe ou d'un groupe à l'autre). Ces objectifs coïncident avec ceux de l'ACP ou de l'ACM.

Ici nous allons également profiter de la structure en groupes sur les variables pour étudier globalement les ressemblances et différences entre groupes : est-ce que les variables d'un groupe apportent la même information que les variables des autres groupes ou bien y a-t-il une information commune et une information spécifique apportée par le groupe, ou bien encore n'y a-t-il aucun lien entre l'information apportée par ce groupe et celle apportée par les autres groupes ?

Pour ce faire, on peut voir si un individu particulier est décrit de la même manière par tous les groupes ou bien si un des groupes le décrit de façon très spécifique. Par exemple, est-ce que le vin numéro 2 est décrit de la même façon par les trois jurys (expert, consommateur et étudiant) ? Si oui, on dira qu'il y a un fort consensus entre les 3 jurys. Est-ce que, au contraire, le vin est décrit de façon spécifique par le jury d'experts par exemple ?

Enfin on pourra comparer la typologie des individus obtenue par chaque groupe. Dans l'exemple des vins cela reviendrait à comparer les trois configurations des vins obtenues grâce à des ACP séparées construites à partir des données de chacun des jurys.

Enfin un point crucial dans l'analyse multi-tableaux est celui de l'équilibre de l'information. L'information d'un groupe ne doit pas écraser, masquer, l'information apportée par les autres groupes et il faut donc veiller à équilibrer l'information apportée par chacun des groupes. Nous verrons dans la prochaine vidéo quel critère est utilisé pour équilibrer les groupes dans l'AFM et nous verrons aussi que cet équilibre peut être géré par une simple pondération des variables.

Deuxième partie. Equilibre et ACP globale

(Diapositives 7 à 14)

Diapositive 7 :

Nous avons vu dans la vidéo précédente que les données sur lesquelles s'applique l'Analyse Factorielle Multiple sont des données où les variables sont structurées en groupes. Nous avons vu qu'il était important d'équilibrer l'influence de chaque groupe de variables dans l'analyse. Voyons maintenant comment équilibrer les groupes et comment mettre en œuvre cet équilibre dans une analyse.

Diapositive 8 :

Tout d'abord, pourquoi est-il important d'équilibrer l'influence de chaque groupe de variables dans l'analyse ?

Notons que nous avons déjà cherché à équilibrer l'influence de variables dans une analyse factorielle. En effet, lorsque l'on construit une ACP normée, on équilibre l'influence de chaque variable pour éviter que les variables qui ont la plus grande variance aient plus d'influence dans l'analyse, *i.e.* plus d'influence dans le calcul des distances entre individus.

De la même façon, avec plusieurs groupes de variables, on veut chercher à équilibrer les groupes de variables pour éviter que certains groupes aient une contribution trop importante dans le calcul des distances entre individus.

Une première idée, qui se rapproche de celle de l'ACP, consiste à équilibrer l'influence de chaque groupe de variables en divisant par l'inertie du groupe de variables. En effet, l'inertie correspond bien à une variance dans un cadre multidimensionnel. On pourrait donc pondérer les groupes de sorte que chaque groupe ait une inertie équivalente. Pour ce faire, il suffit de pondérer chaque variable par l'inertie totale du groupe auquel elle appartient. Avec une telle pondération, si deux groupes n'ont pas le même nombre de variables, le groupe ayant le moins de variables aura la même influence que le groupe ayant plus de variables. De ce point de vue, l'équilibre est bien respecté.

Cependant, cette pondération ne tient pas compte de la méthode utilisée ici : l'analyse factorielle. En effet, l'analyse factorielle est sensible à la répartition de l'inertie d'un groupe à l'autre.

Dans l'exemple de ce petit schéma, un premier groupe est composé de 8 variables fortement corrélées ; le deuxième groupe est composé de 3 variables orthogonales ; enfin le 3ème groupe est identique au second. Avec la pondération par l'inertie totale de chaque groupe, la première dimension du groupe 1 aura un poids proche de 1 (ce groupe est unidimensionnel et toute l'inertie est quasiment sur une seule dimension), tandis que la première dimension des groupes 2 et 3 aura un poids de 1/3 (si l'inertie totale du groupe est de 1, chaque variable représente 1/3 de l'inertie). Par conséquent la première dimension de l'analyse globale coïncidera avec la première dimension du groupe 1. Or on recherche un équilibre pour trouver des structures communes d'un groupe à l'autre donc on aimerait mettre en évidence ici la structure commune aux groupes 2 et 3.

Une deuxième idée consiste à équilibrer les groupes, non plus en fonction de l'inertie totale, mais en fonction de l'inertie maximum d'une dimension. Ainsi, on évite qu'un groupe unidimensionnel ait trop d'importance dans l'analyse mais on ne restreint pas la richesse d'un groupe multidimensionnel, c'est-à-dire d'un groupe qui a plusieurs dimensions de variabilité.

Pour construire un tel équilibre, on divise les valeurs d'une variable par la plus grande valeur propre du groupe auquel elle appartient (on divise par la racine carré de la plus grande valeur propre). Cette pondération permet

dans notre petit exemple de mettre en évidence la structure commune engendrée par les groupes 2 et 3. Ce petit exemple fictif met en évidence une situation relativement fréquente en pratique : des groupes unidimensionnels et d'autres multidimensionnels. L'équilibre proposé évite de donner trop d'importance aux groupes unidimensionnels, groupes qui contiennent une information moins riche.

Diapositive 9 :

Nous avons trouvé ici une pondération qui nous convient, il ne reste plus qu'à rechercher les principales dimensions factorielles comme nous l'avons vu pour l'ACP, l'ACM ou encore l'AFC. La difficulté réside bien dans le choix de la pondération, ce qui a été souligné par JP. Bénzécri qui disait que faire une analyse factorielle, du point de vue des mathématiques, cela revient juste à faire une diagonalisation de matrice, mais que toute la science, tout l'art, était de trouver la bonne matrice à diagonaliser.

La pondération que nous avons trouvée revient à utiliser, pour chaque variable, un poids égal à l'inverse de la première valeur propre du groupe auquel elle appartient. Donc, la première étape de l'analyse consiste à calculer la première valeur propre de chaque groupe. Ensuite, pondérer une variable par l'inverse de l'inertie de la première valeur propre du groupe auquel elle appartient revient à diviser chaque variable par la racine carré de la 1ère valeur propre de son groupe. Dans la notation utilisée ici, *X_j* correspond au tableau *j* centré ou bien centré réduit selon que les variables du groupe *j* sont normées ou non. Enfin, il suffit de faire l'ACP de ce tableau qui juxtapose les variables (pondérées) de tous les tableaux.

Diapositive 10 :

Sur notre jeu de données vin, les ACP construites sur les données de chaque jury donnent les valeurs propres suivantes : la première valeur propre du jury expert, égale à 11.74, est plus grande que les premières valeurs des jurys étudiants et consommateurs. Cela peut s'expliquer par le plus grand nombre de descripteurs évalués par le jury expert.

On peut également noter que les valeurs propres suivantes (les valeurs propres 2 et 3) sont plus grandes pour le jury expert. Si on construisait une ACP sur l'ensemble du jeu de données sans équilibrer préalablement les jurys, le jury expert influencerait fortement les résultats dans le sens où les premières dimensions de l'analyse seraient très liées à la description du jury expert. En équilibrant l'influence de chaque jury grâce à une pondération par la première valeur propre du groupe, chaque jury aura la même contribution dans la construction des dimensions de l'AFM.

Premier point, la pondération choisie est la même pour toutes les variables d'un groupe. Cela permet de préserver la structure du groupe. Autrement, dit, à l'intérieur d'un groupe, l'équilibre entre les variables est préservé : le ratio lambda_2/lambda_1 reste le même, 6.78/11.74=0.58/1. Puisque l'équilibre à l'intérieur d'un groupe reste le même, on peut d'ailleurs réduire ou non les variables d'un groupe selon l'équilibre que l'on souhaite avoir à l'intérieur de ce groupe, cela ne modifie pas l'équilibre entre les groupes.

Plus précisément, grâce à la pondération, la première valeur propre de chaque groupe est ramenée à 1 et donc la principale dimension de variabilité a la même variabilité d'un groupe à l'autre.

Ainsi, un seul groupe ne peut pas générer à lui seul la première dimension de l'AFM, même si ce groupe a beaucoup de variables et est fortement structuré.

Enfin dernier point important, la pondération ramène à 1 la principale dimension de variabilité et les valeurs propres suivantes sont diminuées de façon proportionnelles. Ainsi, un groupe qui est plus multidimensionnel contribue à plus de dimensions car son inertie globale, après équilibre, est plus grande. Un groupe plus multidimensionnel contient plus d'information donc il est normal qu'il contribue à plus de dimensions.

Maintenant est-ce gênant qu'un groupe ait une inertie totale plus grande que les autres? En fait non car cette inertie totale est répartie sur plus de dimensions. On retrouve ici une caractéristique que nous avions déjà vue en analyse des correspondances multiples.

Diapositive 11 :

Nous avons vu que l'AFM peut être considérée comme une ACP sur le tableau pondéré qui juxtapose les données de tous les groupes. Puisque c'est une ACP globale, nous allons obtenir tout d'abord les mêmes graphes qu'en ACP. Un graphe des individus et un graphe des variables. Le graphe des individus permet de visualiser la ressemblance globale entre individus en prenant en compte l'information de toutes les variables. Le graphe des variables permet de visualiser les liaisons entre toutes les variables et bien entendu ces deux graphes sont étudiés simultanément afin de comprendre les différences ou ressemblances entre individus grâce aux variables. Outre les graphes, les indicateurs de qualité de représentation ou de contribution sont également les mêmes qu'en ACP. Et puis, comme pour toute ACP, il est possible d'ajouter des éléments supplémentaires. Donc des individus supplémentaires ou des variables supplémentaires qu'elles soient quantitatives et/ou qualitatives.

Diapositive 12 :

Dans notre exemple voici le graphe des individus. Les vins ont été coloriés selon le cépage : rouge pour les Sauvignon et vert pour les Vouvray. Cet habillage des individus selon la variable cépage met en évidence une opposition entre les deux cépages. Les Sauvignon sont en haut à gauche et les Vouvray en bas à droite. Rappelons que la variable cépage n'était pas active et n'a donc pas été utilisée pour construire les dimensions factorielles. Cette opposition entre cépages est donc bien liée aux variables de description sensorielle évaluée par les juges.

On peut aussi noter que les Vouvray sont plus dispersés dans le plan, ce que l'on peut interpréter de la façon suivante : les Vouvray sont plus différents d'un point de vue sensoriel tandis que les Sauvignon sont plus homogènes.

Enfin, il se dégage plusieurs groupes de vins : par exemple les Vouvray Aubuisières Marigny et Fontainerie Coteau sont très proches sensoriellement, ou encore Fontainerie Domaine et Fontainerie Brûlés sont très proches.

Diapositive 13 :

Le graphe des variables, avec le cercle des corrélations, est le suivant. Les variables ont été coloriées selon le groupe auquel elles appartiennent : en rouge, les variables utilisées par les experts (avec juste le nom du descripteur), en bleu, les variables utilisées par les consommateurs (avec le nom du descripteur suivi de "_C") et en vert les variables utilisées par les étudiants (avec le nom du descripteur suivi de "_E"). Alors bien entendu, ce graphe est très chargé et difficilement lisible car il y a beaucoup de variables. Mais, nous avons plusieurs groupes de variables et nous avons représenté toutes les variables. Donc ce graphe résume beaucoup d'information et c'est pour cette raison qu'il est chargé.

Diapositive 14 :

Mais nous pouvons étudier ce graphe en sélectionnant quelques variables particulières pour la visualisation. Le graphe est donc le même que précédemment et nous avons juste masqué certains libellés, mettant en évidence quelques libellés communs aux différents jurys.

On peut voir par exemple que les variables Odeur de passion sont très liées, ce qui signifie que d'un jury à l'autre cette variable a été comprise de la même façon. Il en est de même pour le sucré. En revanche, la variable acidité a été comprise différemment : les consommateurs et étudiants ont compris cette variable de la même façon tandis que les experts ont compris cette variable différemment. Le coefficient de corrélation entre la variable acidité du jury expert et la variable acidité du jury consommateur est de 0.34 tandis que la corrélation entre la variable acidité du jury étudiant et celle du jury de consommateur est de 0.76. Nous ne pouvons pas savoir qui a compris correctement la variable acidité mais il est possible que les jurys de consommateurs et étudiants aient plus de mal à évaluer l'acidité quand l'équilibre sucré-acide change.

Nous ne rentrons pas plus dans l'interprétation de ce graphe des variables qui est très riche, et qui, combiné avec le graphe des individus, permet d'avoir une description sensorielle des vins obtenue avec les trois jurys d'analyse sensorielle. Ces deux graphes des individus et des variables sont parfaitement équivalents à ceux obtenus en ACP, mais l'AFM bénéficie d'une structure en groupes dont nous n'avons pas encore profité, mis à part dans l'équilibre des groupes. Nous verrons dans la prochaine vidéo l'apport de cette structure en groupes qui est spécifique de l'AFM et qui fait en réalité toute la richesse de l'AFM.

Troisième partie. Etude des groupes

(Diapositives 15 à 27)

Nous avons vu dans la vidéo précédente que l'AFM était une ACP pondérée particulière qui fournissait un graphe des individus et un graphe des variables, parfaitement équivalents aux graphes obtenus en ACP. Voyons maintenant l'apport de la structure en groupes qui est utilisée en AFM.

Diapositive 15 (plan) :

Dans un premier temps, nous verrons comment obtenir une représentation synthétique des groupes de variables qui permettra de comparer globalement l'information apportée par chaque groupe. Ensuite nous utiliserons le cadre commun fourni par l'AFM pour comparer les résultats des ACP construites à partir des données de chaque groupe. Pour cela nous comparerons les nuages d'individus d'une ACP à l'autre puis les dimensions factorielles des ACP séparées.

Diapositive 16 :

Revenons tout d'abord sur la construction des composantes principales. En ACP, la première composante principale est la variable (de norme 1 que j'appelle ici v_1) appartenant à l'espace R^I et qui est la plus liée à l'ensemble des variables, "plus liée" au sens de la covariance au carrée. Si l'ACP est normée, la première composante est la variable la plus liée au sens du coefficient de corrélation au carré.

Puisque l'AFM est une ACP sur un tableau pondéré, nous pouvons écrire le critère précédent en faisant apparaître la pondération utilisée par l'AFM. Chaque variable *x_k* est divisée par la racine carrée de la 1ère valeur propre du groupe auquel elle appartient. On doit sommer sur toutes les variables mais comme les variables sont organisées par groupe, on peut écrire le critère en faisant apparaître une somme sur les groupes et une somme sur toutes les variables du groupe j.

On peut sortir la première valeur propre lambda_1^j et mettre en évidence que la première composante de l'AFM est la variable (de norme 1) qui maximise la covariance au carré avec les variables de tous les groupes, la covariance étant divisée par l'inertie axiale maximum de chaque groupe. L' "inertie axiale maximum" correspond à l'inertie de la dimension la plus forte du groupe.

Si on note Lg(Kj,v1) le terme en rouge,

ce coefficient Lg(Kj,v1) correspond à l'inertie projetée de toutes les variables du groupe j sur la variable v1. C'est une mesure de liaison entre un groupe de variables et une variable. Il existe une autre mesure de liaison bien connue entre une variable et un groupe de variables, il s'agit du coefficient de détermination qui est utilisé en régression. L'intérêt du Lg est que ce critère est basé sur l'inertie et en analyse factorielle on essaie toujours de maximiser une inertie. La formule nous dit donc que la première composante de l'AFM est la variable qui maximise la liaison avec tous les groupes au sens du Lg. Pour trouver la 2ème composante principale de l'AFM, il faut chercher, parmi les variables orthogonales à la variable v1, la variable qui est la plus liée à l'ensemble des groupes selon le critère du Lg. Et ainsi de suite pour les composantes principales suivantes.

Le coefficient Lg varie entre 0 et 1. Il vaut 0 si toutes les variables du groupe j sont non-corrélées avec la variable v1. Il vaut 1 si la variable v_1 est confondue avec la première composante principale du groupe. En effet, la première composante principale du groupe est par définition la variable la plus liée à toutes les variables du groupe et maximise donc la somme des covariances au carré. Cette somme des covariances est égale à l'inertie

de la première composante principale qui vaut la première valeur propre du groupe. Comme les variables sont pondérées par la première valeur propre du groupe, le Lg maximum correspond bien à la valeur 1.

Diapositive 17 :

La mesure de liaison Lg que nous venons de définir permet d'appréhender la liaison entre une variable et un groupe de variables. On peut donc utiliser cette mesure pour voir la liaison entre un groupe de variables et une composante principale. Cela nous suggère d'utiliser cette mesure pour construire un graphe des groupes. Un groupe aura pour abscisse le coefficient Lg entre ce groupe et la première dimension et comme ordonnée le coefficient Lg avec la deuxième dimension.

Dans notre exemple, la représentation des groupes par rapport aux composantes principales de l'AFM est la suivante.

On peut voir que tous les jurys ont une abscisse proche de 1 et donc un Lg élevé avec la première dimension principale de l'AFM. Or un Lg élevé signifie que la composante principale de l'AFM est liée à une dimension forte du groupe (la dimension principale ou une dimension presque aussi importante que la dimension principale). Ainsi, dans notre exemple, la première dimension de l'AFM est présente dans tous les groupes et c'est donc une dimension commune à tous les groupes.

On peut voir que la deuxième dimension est beaucoup plus liée au groupe expert qu'aux autres groupes. Ceci n'est pas vraiment surprenant car nous avions vu que le groupe expert avait une 2ème valeur propre plus grande (relativement à sa 1ère valeur propre) que pour les 2 autres groupes.

Enfin les points représentant les groupes étudiants et consommateurs sont proches ce qui signifie que ces groupes ont en commun plusieurs dimensions et qu'ils induisent donc à peu près la même structure sur les individus.

Cette représentation graphique fournit une représentation synthétique des groupes et il est très facile de voir quels groupes se ressemblent globalement. Deux groupes se ressemblent s'ils sont proches sur toutes les dimensions, mais comme les premières dimensions résument l'essentiel de l'information, il est souvent suffisant de se contenter de comparer les groupes sur les premières dimensions (3 ou 4 suffisent généralement).

Ce graphique permet de comparer globalement les groupes et donc de voir si les distances entre individus sont similaires d'un groupe à l'autre. Autrement dit, si les nuages de points vus par un groupe de variables ou par un autre sont similaires.

Diapositive 18 :

Nous avons défini la mesure Lg entre une composante et un groupe de variables, mais cette mesure Lg peut être étendue très facilement à la liaison entre 2 groupes de variables. Il suffit de calculer la somme de toutes les covariances entre les variables d'un groupe et les variables de l'autre; les variables étant bien entendu pondérées par la pondération de l'AFM.

Si on calcule le coefficient Lg entre un groupe et lui même, cela revient à calculer la dimensionnalité de ce groupe. En réécrivant le critère, on fait apparaître les ratios entre les valeurs propres au carré de chaque dimension divisées par la première valeur propre au carré du groupe. Si la 1ère valeur propre est beaucoup plus grande que les autres, alors le Lg sera proche de 1 et le groupe sera presque unidimensionnel. Au contraire, si beaucoup de valeurs propres sont proches de la 1ère valeur propre du groupe, alors plusieurs dimensions de variabilité sont importantes et le groupe sera multidimensionnel. Ce coefficient peut donc être vu comme un indice de dimensionnalité d'un groupe.

Enfin, l'inconvénient du coefficient Lg entre 2 groupes est que c'est un critère qui n'est pas borné. Si on normalise le coefficient Lg entre 2 groupes par la dimensionnalité de chacun des groupes, on définit le coefficient RV. Ce coefficient RV varie entre 0 et 1 et il est plus facile de savoir si 2 groupes sont liés ou non puisque le coefficient est borné par 1. Notons toutefois que ce coefficient a tendance à être plus élevé quand le nombre d'individus est petit ou quand le nombre de variables dans chaque groupe est grand.

Diapositive 19 :

Dans notre exemple, les coefficients Lg puis les coefficients RV entre les groupes sont les suivants. La matrice étant symétrique, nous donnons ici uniquement les termes sous la diagonale.

A partir du Lg, on peut voir que le jury expert donne une description plus multidimensionnelle, et donc plus riche, des produits que les autres jurys car le Lg est plus élevé.

Si on regarde les coefficients RV, on voit que les jurys étudiant et expert sont proches car le RV est proche de 1 (0.85).

Enfin, on peut considérer la configuration moyenne de l'AFM et voir la liaison entre la configuration moyenne de l'AFM et chaque groupe. La configuration moyenne de l'AFM est ici un groupe correspondant à l'ensemble des coordonnées des individus sur toutes les dimensions factorielles; il s'agit de la configuration commune. Ces coefficients RV sont donnés dans la ligne MFA. On peut voir que le jury étudiant a une configuration très proche de la configuration moyenne puisque le RV est très proche de 1 (il vaut 0.96).

Diapositive 20 :

Voyons maintenant une autre façon de comparer les groupes à partir des configurations des individus de chaque groupe. Pour cela, les typologies fournies par chaque groupe vont être comparées dans un cadre commun, plus précisément dans le référentiel commun obtenu grâce à l'AFM. Comparer les typologies revient à voir si les individus sont vus de la même façon par les différents groupes ou bien si certains individus sont particuliers pour quelques groupes.

Diapositive 21 :

Pour ce faire, nous allons procéder comme suit. Illustrons la démarche avec un jeu de données comportant 3 groupes de variables. L'AFM est tout d'abord construite à partir de ces 3 groupes comme nous l'avons vu. Nous avons alors une représentation des individus sur les principales dimensions de l'AFM. Les individus sont dans un espace R^K à K dimensions, cet espace est la somme directe des espaces R^Kj engendrés par chaque groupe de variables. Essayons maintenant de représenter chaque individu du jeu de données uniquement à partir des données d'un seul groupe de variables. Appelons individu partiel i exposant j l'individu i vu par les seules variables du groupe j. Comment représenter cet individu partiel ? Et surtout comment représenter tous les individus partiels associés à l'individu i (à savoir les individus partiels i^1, i^2, jusqu'à i^J) dans un même référentiel ?

Le référentiel commun est l'espace fourni par l'AFM à partir des données de tous les groupes. On juxtapose alors au tableau de données, un tableau par groupe. Dans le sous-tableau permettant la projection du groupe j, toutes les variables qui n'appartiennent pas au groupe j prennent des valeurs nulles (des valeurs nulles car le tableau est centré, sinon les valeurs prises sont la moyenne de la variable). Et les valeurs des variables appartenant au groupe j sont les variables centrées (éventuellement réduites si on réduit les variables de ce

groupe), et divisée par la racine carrée de la première valeur propre du groupe. Autrement dit, les valeurs utilisées dans l'analyse globale de l'AFM. Ces individus partiels sont alors utilisés comme individus supplémentaires et donc projetés sur les dimensions de l'analyse globale.

On a alors la projection de l'individu i vu par les variables du groupe 1, ici en rouge. Une remarque importante cependant : comme nous avons des 0 pour tous les groupes autres que le groupe 1, alors le point partiel du groupe 1 a tendance à se rapprocher du centre du nuage. Ceci est vrai pour tous les individus partiels de chacun des groupes. Pour s'en convaincre, prenons l'exemple trivial où les trois groupes de variables sont identiques. Tous les points partiels d'un même individu devraient être superposés et superposés à l'individu moyen. Or, si on calcule la projection d'un individu partiel d'un groupe, comme il prend des valeurs nulles sur les 2 autres groupes, sa coordonnée sera proche du barycentre du nuage. Elle sera même 3 fois trop proche du barycentre. Ainsi, la projection de tous les points partiels doit être dilatée de 3 dans notre exemple. Quand le jeu de données comporte J groupes, les points partiels doivent être projetés comme des points supplémentaires puis dilatés d'une constante J.

Avec cette dilatation, nous avons pour chaque individu autant de points partiels qu'il y a de groupes, et le point moyen, correspondant à l'individu vu dans l'analyse globale, est au barycentre des points partiels.

Diapositive 22 :

La représentation des points partiels est donc très utile puisqu'elle permet de comparer la représentation des individus d'un groupe à l'autre. Prenons l'exemple suivant. Des individus ont tout d'abord été interrogés sur plusieurs questions relatives à leurs opinions. Ensuite ils ont été interrogés sur plusieurs questions relatives à leur comportement. Ces deux séries de questions peuvent être considérées comme deux groupes de variables dans l'AFM. L'AFM fournit alors une représentation des individus vus par l'ensemble des variables, il s'agit du point correspondant à l'analyse globale, et donc du point moyen ici en violet. L'interprétation des positions des points moyens sur les axes peut se faire comme pour une ACP. On peut représenter également les points partiels correspondant à l'individu vu uniquement à partir des variables d'opinion puis vu uniquement à partir des variables comportementales. Si les deux points partiels sont proches, et donc proches du point moyen, alors l'individu i' a un comportement en accord avec ses opinions, tandis que si les points partiels sont éloignés, alors l'individu i a un comportement qui n'est pas en accord avec ses opinions. L'écart entre les 2 points partiels est une visualisation de la discordance comportementale.

Prenons un second exemple avec un questionnaire auquel des participants à un cours ont répondu avant de commencer le cours. Plusieurs questions sont relatives à leurs attentes vis-à-vis du cours. Un second questionnaire a été envoyé aux mêmes personnes à l'issue du cours. Les questions sont alors relatives à ce que les participants ont appris pendant le cours. L'individu à droite a ses deux points partiels proches, tandis que l'individu à gauche a ses points partiels éloignés.

L'individu de droite a donc trouvé dans le cours ce qu'il attendait du cours, on peut donc dire qu'il est satisfait. L'individu de gauche n'a en revanche pas appris ce qu'il attendait du cours. On peut dire qu'il a été surpris. Mais a-t-il été surpris dans le mauvais sens, autrement dit a-t-il été déçu, ou bien a-t-il été surpris dans le bon sens ? Pour le savoir, il suffit de revenir à l'interprétation des axes, et plus spécialement ici à l'interprétation de l'axe 1 qui sépare les 2 points partiels. Si les coordonnées positives sur l'axe sont relatives à une forte appréciation, alors l'individu a appris plus que ce qu'il espérait du cours et on pourra dire qu'il a été agréablement surpris, dans le cas contraire on dira qu'il a été déçu. Ceci montre que la position des points moyens, i.e. des individus vus par l'ensemble des variables, et la position des points partiels peuvent s'interpréter à partir d'une seule représentation sur les facteurs communs de l'AFM.

Diapositive 23 :

Les relations de transition que nous avons vues en ACP peuvent s'appliquer directement sur les points moyens de l'AFM. Rappelons que ces relations de transition permettent d'interpréter la position d'un individu grâce aux graphes des variables. Un individu est du côté des variables pour lequel il prend de fortes valeurs et à l'opposé des variables pour lequel il prend de faibles valeurs.

Ces relations de transition sont aussi adaptées pour les points partiels. Il s'agit juste de la restriction de la relation précédente aux variables du groupe j. Comme nous l'avons précisé précédemment, nous multiplions les coordonnées des points partiels par J (le nombre de groupes) pour que les coordonnées des points partiels et des points moyens puissent être interprétées dans un cadre unique et dans un même repère. La dilatation par J appliquée à tous les points partiels permet que le point moyen soit bien au barycentre de ses points partiels et que points moyens et partiels soient représentés dans un même repère.

Diapositive 24 :

Dans notre exemple des vins, la représentation superposée est la suivante. Les points en noir correspondent aux vins vus par l'ensemble des jurys. Les points en couleur correspondent aux points partiels : en rouge on trouve les vins vus uniquement par le jury expert, en bleu vus uniquement par le jury de consommateurs et en vert vus uniquement par le jury étudiant.

Chaque point noir correspondant au vin moyen est bien au barycentre des points de couleur.

Tous les vins sont vus de façon relativement homogène par les trois jurys sensoriels car tous les points partiels relatifs à un même vin sont proches du point moyen. On peut noter cependant que le vin Aubuisières Silex, en bas à gauche, a été perçu plus extrême par les consommateurs que par les experts pour les variables liées à la première dimension. En effet, l'abscisse de ce vin est fortement négative pour les consommateurs, et relativement proche de 0 pour les experts.

Diapositive 25 :

Visuellement on voit que les coordonnées des points partiels sont relativement similaires d'un groupe à l'autre pour les deux premières dimensions. On peut définir un indicateur pour mesurer rapidement et globalement si les coordonnées des points partiels sont proches de celles des points moyens dimension par dimension.

Pour cela, pour une dimension, on va décomposer l'inertie totale des points partiels. Pour la dimension s, l'inertie totale des points partiels peut se décomposer en l'inertie inter points moyens (i.e. inter individus) plus l'inertie intra individu. Et on peut calculer le ratio inertie inter sur inertie totale.

Dans l'exemple, pour la dimension 1, le ratio vaut 0.93 et est donc très proche de 1, ce qui signifie que les coordonnées des points partiels relatives à un même individu sont très proches entre elles et donc très proches des coordonnées de leur point moyen correspondant.

L'inertie intra, qui mesure la ressemblance entre nuages partiels dimension par dimension, peut aussi être décomposée par individu. Cette décomposition par individu est très utile pour trier les individus par inertie intra décroissante et identifier quels individus sont vus de façon différente par chaque groupe de variables. Lors de l'interprétation des différences entre groupes, on pourra s'appuyer sur ces individus.

Diapositive 26 :

Nous avons comparé les groupes de variables grâce au graphe des points partiels qui représente les typologies des individus de chacun des groupes dans un espace commun. Il est également possible de comparer les

groupes à partir des dimensions factorielles des analyses séparées, autrement dit en comparant les composantes principales de l'ACP de chacun des groupes.

Pour cela, nous allons considérer les composantes principales de l'ACP d'un groupe.

Et nous allons les utiliser comme variables supplémentaires dans l'analyse globale, donc dans l'AFM. Cela nous permettra de projeter les composantes principales de ce groupe et de voir comment ces composantes sont liées aux dimensions de l'AFM. Si les composantes principales d'un groupe sont très corrélées aux dimensions factorielles de l'AFM, alors le graphe des individus de l'ACP de ce groupe a la même forme que le graphe des individus des points moyens de l'AFM.

Diapositive 27 :

Dans notre exemple, voici le graphe qui donne la projection des composantes principales des ACP séparées sur les deux premières dimensions de l'AFM.

La première composante de l'ACP des étudiants est extrêmement corrélée avec la première composante de l'AFM. Ainsi les vins qui ont une faible coordonnée (resp. une forte coordonnée) sur l'axe des abscisses de l'ACP sur les données étudiants ont une faible coordonnée (resp. une forte coordonnée) sur l'axe des abscisses de l'AFM. Et il en est de même pour la 2ème composante de l'ACP des étudiants qui est également extrêmement corrélée avec la 2ème composante de l'AFM. Cela signifie que la typologie des vins fournie par les étudiants est extrêmement proche de la typologie des vins obtenue par l'AFM et donc par l'ensemble des jurys.

Pour les experts les composantes 1 et 2 de l'ACP ne coïncident pas exactement avec les composantes 1 et 2 de l'AFM. Cependant, les composantes de l'ACP sont toutes les deux bien projetées ce qui signifie que le plan 1-2 de l'ACP construite sur les données expert est proche du plan 1-2 de l'AFM mais à une rotation près. Il en est de même pour les consommateurs.

Quatrième partie. Compléments

(Diapositives 28 à 39)

Nous avons vu dans les vidéos précédentes comment prendre en compte la structure en groupes avec une analyse factorielle multiple quand tous les groupes contiennent des variables quantitatives.

Diapositive 28 :

Nous allons voir maintenant quelques compléments en commençant par voir comment prendre en compte des groupes de variables qualitatives. Ensuite nous verrons comment faire quand un ou plusieurs groupes de variables correspondent à un ou plusieurs tableaux de contingence. Et enfin nous verrons les aides à l'interprétation utiles pour interpréter les résultats d'une AFM. Nous insisterons sur les aides spécifiques de l'AFM.

Diapositive 29 :

Considérons un tableau de données où toutes les variables sont qualitatives et où les variables sont structurées par groupe. La problématique est alors similaire à celle où les variables sont toutes quantitatives. En effet, on va chercher à équilibrer l'influence de chaque groupe de variables dans l'analyse globale, puis on va construire des graphes spécifiques de l'AFM pour étudier les similitudes et différence entre groupes. Le graphe des individus sera identique à ceux de l'ACM et donc il contiendra également les modalités. Une modalité étant au barycentre des individus qui la prenne. Ensuite, on pourra construire le graphe des représentations superposées, le graphe des groupes et le graphe des axes partiels pour voir les relations entre l'analyse globale et les analyses séparées.

La démarche est donc la même que précédemment sauf que, sur chaque groupe de variables, on utilise l'ACM plutôt que l'ACP. Ainsi, pour chaque groupe, on construira un tableau disjonctif complet, puis on pondèrera comme en ACM en fonction de la marge colonne. Il reste à équilibrer les groupes en pondérant par la première valeur propre de l'ACM du groupe.

Diapositive 30 :

Dans l'exemple, on peut ajouter le groupe de variables qualitatives restreint à la seule variable de cépage. Ceci permet d'illustrer les sorties obtenues quand certains groupes sont qualitatifs, que ces groupes soient actifs ou supplémentaires. Nous avons ici le graphe des groupes où on voit que le groupe cépage a une coordonnée proche de 0.4 sur les deux axes. Cela signifie que ce groupe est lié aux 2 premières dimensions de l'AFM et donc que les vins des deux cépages sont séparés sur les deux dimensions.

C'est ce que l'on peut vérifier sur le graphe des individus en coloriant les vins selon la variable qualitative cépage.

Le graphe des axes partiels montre une seule dimension pour le groupe qualitatif. Ceci est normal car comme il n'y a qu'une seule variable à deux modalités, l'espace engendré est de dimension 1 (le nombre total de modalités, ici 2, moins le nombre total de variables, ici 1). Cette dimension est bien projetée sur le plan de l'AFM donc la variabilité induite par le groupe de variables qualitatives est bien récupérée par le plan de l'AFM.

Enfin, on peut construire le graphe des points partiels des individus et des modalités. Ici, nous avons construit les points partiels uniquement des modalités pour éviter de surcharger le graphe. On voit que le jury expert a tendance à beaucoup plus séparer les Vouvray des Sauvignon car les points partiels sont beaucoup plus extrêmes comparés aux points partiels des consommateurs qui sont proches du centre du graphe.

Diapositive 31 :

Il est aussi possible de mixer la nature des variables d'un groupe à l'autre et étudier par une AFM des tableaux de données dans lesquels certains sous-tableaux sont constitués de variables quantitatives quand d'autres sont constitués de variables qualitatives. A l'intérieur de chaque groupe, les liaisons entre variables sont étudiées comme pour une ACP quand les variables sont quantitatives et comme pour une ACM quand les variables sont qualitatives.

On équilibre ensuite l'influence de tous les groupes en utilisant la pondération de l'AFM par la première valeur propre de l'analyse factorielle construite sur les données d'un groupe (première valeur propre de l'ACP ou première valeur propre de l'ACM selon la nature des variables du groupe). Les représentations graphiques seront les mêmes que celles que nous avons déjà présentées : graphe des individus et des modalités, graphe des variables, puis représentations spécifiques avec le graphe des groupes, la représentation des points partiels et la représentation des axes partiels.

Notons le cas particulier suivant : si chaque groupe de variables n'est constitué que d'une seule variable, quantitative ou qualitative, alors on tombe sur une méthode appelée l'analyse factorielle de données mixtes. Cette méthode permet d'analyser des tableaux simples constitués de variables quantitatives et qualitatives. Les représentations sont un mixte entre celle de l'ACP (pour les variables quantitatives) et celle de l'ACM avec un graphe des individus et des modalités et un graphe de variables avec le carré des liaisons (carré des rapports de corrélation pour les variables qualitatives et carré des corrélations pour les variables quantitatives). Cette analyse factorielle multiple de données mixtes équilibre donc l'influence de toutes les variables quantitatives et qualitatives. Elle est souvent utile car de nombreux tableaux de données contiennent simultanément des variables quantitatives et qualitatives.

Diapositive 32 :

L'AFM peut aussi être étendue aux tableaux de contingence. Ainsi certains groupes de colonnes peuvent être des tableaux de contingence. Tous ces tableaux doivent alors avoir une même dimension commune.

La principale difficulté est alors de trouver une pondération pour les lignes du tableau puisque le poids des lignes est différent d'un tableau de contingence à l'autre (les poids seraient identiques dans le cas où la somme des effectifs d'une ligne reste le même d'un tableau à l'autre, ce qui arrive très peu souvent en pratique). On choisira comme poids commun d'une ligne la somme des effectifs de cette ligne sur tous les tableaux de contingence divisée par la somme totale des effectifs de tous les tableaux de contingence. L'analyse des résultats et les principales interprétations d'une AFM sur tableaux de contingence sont exactement les mêmes que ceux d'une AFM classique.

Donnons quelques exemples d'applications. Tout d'abord, on trouve de nombreux exemples en analyse textuelle ce qui correspond à une extension des tableaux de données textuelles analysés par AFC. Citons l'exemple suivant où une même enquête a été effectuée dans plusieurs pays. Chaque colonne correspond à une question, les lignes correspondent aux modalités d'une variable qualitative (les classes d'âge par exemple) et dans une cellule on trouve le nombre de répondants ayant répondu oui à une question. Chaque pays correspond à un tableau et on va chercher à voir l'information commune à tous les pays et l'information spécifique de chaque pays grâce à l'AFM.

On peut aussi avoir des exemples en écologie, avec un tableau croisant site et espèce recueilli sur plusieurs années. On cherchera alors à voir l'évolution des associations entre site et espèce au cours du temps. Les sites, en lignes, peuvent être les mêmes d'une année à l'autre tandis que certaines espèces peuvent disparaître et ne plus être présentes certaines années.

L'AFM peut aussi gérer simultanément des tableaux de fréquences, des tableaux de variables quantitatives et des tableaux de variables qualitatives. Le poids des lignes est alors calculé à partir des tableaux de fréquences actifs.

Diapositive 33 :

On peut maintenant revoir quelques aides à l'interprétation communes à toutes les analyses factorielles, ACP, ACM, AFM.

Tout d'abord, il est possible d'ajouter des informations supplémentaires, individus supplémentaires ou variables supplémentaires et bien entendu en AFM des groupes de variables supplémentaires. Ces éléments supplémentaires ne participent pas à la construction des axes mais pourront aider à l'interprétation.

Dans notre exemple sur les vins, nous disposions de 3 groupes de variables correspondants aux dégustations de 3 jurys sensoriels, mais nous disposions aussi de données de préférences. Un nouveau groupe de consommateurs a dégusté les vins en mettant une note d'appréciation (0 pour je n'aime pas du tout ce vin et 10 pour j'apprécie beaucoup ce vin). On peut alors constituer un tableau de données avec en lignes les 10 vins et en colonnes les 60 dégustateurs et, dans une case du tableau, la note attribuée pour un vin par un dégustateur. On juxtapose alors ce tableau au tableau de données utilisé dans l'AFM puisque les lignes sont bien les mêmes. On peut aussi utiliser la variable cépage comme information supplémentaire mais nous avons déjà vu comment analyser les résultats de ce groupe qualitatif donc nous nous contenterons d'analyser les résultats du groupe supplémentaire quantitatif.

On peut alors se poser la question suivante : est-ce que l'appréciation des vins est liée à la description sensorielle ? Pour répondre à cette question, nous allons refaire une AFM en considérant le groupe de préférence, constitué de 60 variables en supplémentaire.

Diapositive 34 :

Le groupe de préférences ayant été mis en supplémentaire, le graphe des individus moyens et le graphe des variables obtenus par l'AFM sont les mêmes que précédemment. Ils donnent la description sensorielle des vins construite à partir des dégustations des 3 jurys.

Le groupe de préférence a des coordonnées relativement fortes sur les deux premiers axes de l'AFM. Cela signifie que globalement ce groupe est lié aux deux axes de l'AFM et donc à la description sensorielle des vins. L'appréciation des vins n'est pas indépendante de la description sensorielle.

Plus dans le détail, toutes les variables de préférences peuvent être représentées sur le graphe des variables, mais par souci de lisibilité, nous les avons représentées sur un second graphe. Chaque flèche correspond à un dégustateur et on voit que la plupart des flèches sont dirigées en bas à gauche et sont plutôt bien représentées. Cela signifie que le vin en bas à gauche est fortement apprécié par beaucoup de dégustateurs. C'est le vin Aubuisières Silex, qui est un vin qui a été décrit comme sucré par les 3 jurys, comme on peut le voir sur le graphe des variables sensorielles.

Diapositive 35 :

On retrouve en AFM les indicateurs classiques de contribution et de qualité de représentation. Pour les individus et les variables, ces indicateurs sont calculés exactement comme en ACP.

Pour les groupes, nous pouvons aussi calculer ces indicateurs. Tout d'abord, la contribution d'un groupe est égale à sa coordonnée (et non sa coordonnée au carrée), divisée par la somme des coordonnées si on veut

exprimer la contribution relative. On peut donc lire directement la contribution d'un groupe à la construction d'un axe à partir du graphe des groupes.

Enfin, la qualité de représentation d'un groupe est mesurée comme usuellement par le cosinus carré de l'angle entre le vecteur partant du centre du nuage et allant jusqu'au point représentant le groupe dans l'espace et le vecteur partant du centre du nuage et allant sur le projeté du groupe sur l'axe.

On peut alors sommer les cosinus carrés sur plusieurs axes pour avoir la qualité de représentation dans un plan ou un sous-espace.

Diapositive 36 :

Comme pour les autres méthodes d'analyse factorielle, on peut décrire les dimensions factorielles de façon automatique en calculant les liaisons entre chaque variable et chaque dimension. Ceci est particulièrement utile en AFM car le nombre de variables est souvent important. Pour décrire les dimensions par les variables quantitatives, on va calculer le coefficient de corrélation entre chaque variable et la dimension puis trier les coefficients de corrélation par valeur décroissante, et ne conserver que les coefficients significativement différents de 0.

Dans l'exemple, nous avons utilisé ici uniquement les variables actives, c'est-à-dire les descripteurs sensoriels pour décrire les dimensions. Nous aurions pu également décrire les dimensions par les variables supplémentaires mais ici les préférences ne sont pas intéressantes pour comprendre les axes. La variable la plus liée à la première dimension est liée négativement à la première dimension et il s'agit de la typicité du goût Chenin évalué par les consommateurs. Ensuite, c'est l'odeur de vanille évaluée par les experts, variable corrélée positivement à la première dimension, qui permet de décrire le premier axe. Le 2ème axe est quant à lui lié (négativement) à l'odeur végétale évaluée par les consommateurs. Nous avons ainsi une aide pour voir parmi les 57 variables quelles sont celles qui expliquent les dimensions.

Diapositive 37 :

Pour décrire les axes par les variables qualitatives, on construit un modèle d'analyse de variance à 1 facteur dans lequel on explique les coordonnées des individus (des vins) sur l'axe en fonction de la variable qualitative. On trie alors les variables qualitatives en fonction des rapports de corrélation et on conserve les variables qualitatives dont le rapport de corrélation est significativement différent de 0. Cela permet de voir globalement quelles variables sont les plus liées aux dimensions. Par ailleurs, on construit des tests de Student pour trier les modalités. On va tester si la coordonnée d'une modalité est significativement différente de 0 ou non. Là encore, les modalités sont triées en fonction de la probabilité critique du test et en fonction du signe du coefficient.

Dans notre exemple, la variable cépage est liée aux deux premières dimensions. La probabilité est très légèrement inférieure à 5% et donc on rejette l'hypothèse que le rapport de corrélation est égal à 0. Rappelons qu'ici la variable cépage est supplémentaire et n'a donc pas été utilisée pour construire les axes. Ainsi, le test construit est bon. En revanche, pour les variables sensorielles, les tests doivent être utilisés avec plus de prudence puisque les variables ont été utilisées pour construire les dimensions et sont donc mécaniquement liées aux axes. Si on regarde les modalités, on voit que les vins Vouvray ont des coordonnées significativement positives sur le premier axe et significativement négatives sur le deuxième axe. Ils sont donc en bas à droite tandis que les Sauvignon sont en haut à gauche.

Diapositive 38 :

Pour terminer, voici un transparent récapitulant les étapes importantes pour mettre en œuvre une AFM.

Tout d'abord, la première question à se poser est : y a-t-il une structure en groupes dans le jeu de données ou bien doit-on faire une analyse d'un tableau simple avec une ACP (si les données sont quantitatives) ou une ACM (si les données sont qualitatives) ? S'il y a une structure en groupes, quels groupes construire ? Un groupe correspond à un sous-tableau et donc le plus souvent à un type d'information. Cette constitution des groupes est essentielle puisqu'elle définit la pondération et l'équilibre entre groupes.

Ensuite, il faut choisir les groupes qui seront actifs et ceux qui seront supplémentaires.

Pour les groupes de variables quantitatives, il faut déterminer si on veut que les variables du groupe soient réduites ou non-réduites selon l'importance que l'on veut donner aux variables à l'intérieur du groupe. Il s'agit bien de voir l'importance des variables à l'intérieur d'un groupe et non d'un groupe à l'autre car la pondération de l'AFM par la première valeur propre de chaque groupe conduit à accorder la même importance à chaque groupe de variables.

Ensuite l'AFM est réalisée.

Il faut alors choisir le nombre d'axes à interpréter en regardant par exemple un graphe des valeurs propres de l'AFM.

Ensuite on commence par interpréter conjointement le graphe des individus et le graphe des variables avant d'utiliser les sorties spécifiques de l'AFM pour étudier globalement les groupes et voir quels sont les groupes qui se ressemblent grâce au graphe des groupes, quels sont les individus qui ne sont pas vus de façon homogène par chaque groupe grâce au graphe des points partiels.

On peut ensuite visualiser les analyses séparées pour voir quelles sont les dimensions de chaque groupe qui sont liées aux dimensions de l'AFM.

Enfin, on utilise les indicateurs, qualité de représentation, contribution, coefficient Lg ou RV pour enrichir l'interprétation.

Quelques logiciels ont une fonction permettant de faire une AFM. La fonction la plus avancée est la fonction MFA du package FactoMineR puisqu'elle permet de faire des AFM avec tout type de groupes: des groupes de variables quantitatives ou des groupes de variables qualitatives ou encore des tableaux de contingence. Chaque groupe peut être soit actif soit supplémentaire. Vous retrouverez également toutes les aides à l'interprétation décrites dans ce cours.

Diapositive 39 :

Pour conclure, l'AFM est une méthode multi-tableaux utile pour analyser des tableaux ayant les mêmes lignes. Les groupes ou tableaux doivent avoir des variables quantitatives, qualitatives ou bien correspondre à des tableaux de fréquences.

L'intérêt de l'AFM est d'équilibrer l'influence de chaque groupe puis de représenter l'information apportée par tous les groupes dans un référentiel commun. L'AFM fournit d'une part deux graphes, un graphe des individus et un graphe des variables, qui sont des sorties classiques en analyse factorielle. Ces graphes s'analysent comme en ACP ou en ACM. Mais l'AFM fournit également des sorties spécifiques et c'est ce qui en fait sa richesse. Les sorties spécifiques permettent de comparer l'information apportée par chaque tableau : l'information peut être comparée de façon très globale grâce au graphe des groupes.

On voit alors si globalement les groupes ont les mêmes dimensions, et si les dimensions mises en avant par l'AFM sont présentes dans chaque groupe. L'information des groupes peut être comparée de façon plus précise

grâce à la représentation des dimensions des analyses séparées : en effet, on pourra voir si les dimensions de variabilité d'un tableau (obtenues par une ACP, une ACM ou une AFC) sont liées ou non aux dimensions communes, i.e. aux dimensions de l'AFM. Enfin l'information apportée par chaque groupe peut aussi être comparée au niveau de chaque individu grâce au graphe des points partiels.

Voici deux livres sur l'analyse factorielle multiple écrits par Jérôme Pagès. La dernière référence est dédiée à l'analyse factorielle multiple. Vous pourrez trouver dans ce livre plus de détails sur les justifications mathématiques des méthodes mais également quelques exemples d'utilisation de l'AFM.

Vous avez vu l'ensemble du cours sur l'analyse factorielle multiple, vous pouvez voir maintenant comment mettre en oeuvre l'analyse factorielle multiple grâce à FactoMineR. N'oubliez pas non plus de faire les quiz et exercices associés à ce cours.