Transcription de l'audio du cours d'Analyse des correspondances multiples

Première partie. **Données - problématique**
Diapositives 1 à 9
Pages 2 à 5

Deuxième partie. **Visualisation du nuage des individus**
Diapositives 10 à 23
Pages 6 à 10

Troisième partie. **Visualisation du nuage des modalités et représentation simultanée**
Diapositives 24 à 29
Pages 11 à 14

Quatrième partie. **Aide à l'interprétation**
Diapositives 30 à 39
Pages 15 à 18

Première partie. Données - problématique

(Diapositives 1 à 9)

Diapositive 1

Cette semaine, nous vous présentons 4 vidéos de cours sur l'analyse des correspondances multiples. Les principales caractéristiques de la méthode sont décrites à partir d'un exemple d'application.

Diapositive 2 (plan)

L'ensemble des vidéos aborde les points suivants : nous commencerons, dans cette vidéo, par décrire les données sur lesquelles on travaille en ACM. A partir de ces données nous dégagerons des objectifs et une problématique. Ceci va nous conduire à une transformation du tableau de données. En analyse des correspondances multiples, comme dans toute analyse factorielle, on construit des nuages de points, nuage des lignes et nuage de colonnes. Ici il s'agit d'un nuage des individus et d'un nuage des modalités. Nous verrons comment visualiser le nuage des individus et comment l'interpréter grâce aux modalités. Nous verrons ensuite comment visualiser directement le nuage des modalités. Nous montrerons qu'en ACM le nuage des individus et celui des modalités peuvent être représentés simultanément sur un même graphe, c'est ce qu'on appelle la représentation simultanée des deux nuages. Nous verrons enfin dans une dernière vidéo les différentes aides à l'interprétation classique en analyse factorielle. Tout ceci va être illustré à partir d'une enquête de l'INSEE sur les loisirs des Français.

Diapositive 3

Les données sur lesquelles nous travaillons sont constituées par un tableau rectangulaire avec *I* individus en lignes et *J* variables qualitatives en colonnes. Attention distinguons bien une variable qualitative par exemple la couleur, des modalités qu'elle peut prendre par exemple bleu. A l'intersection de la ligne i et de la colonne *j* on va trouver V_{ij} la modalité de la variable j possédée par l'individu *i*.

Exemple. L'exemple classique en ACM est celui d'une enquête. On a interrogé *I* personnes et le questionnaire comprend *J* questions à choix multiples. Exemple de question à choix multiples : quelle est votre catégorie socioprofessionnelle ? Les modalités de réponse possibles sont ouvriers, employés, etc. Autre exemple de question : quelle est votre situation de famille? célibataire, marié, etc. La situation de famille, c'est la variable; célibataire, marié, ce sont les modalités que peut prendre cette variable. C'est ce tableau que l'on construit et que l'on fournit au logiciel d'analyse des correspondances multiples.

Diapositive 4

A partir de ce tableau brut, on construit un tableau dit tableau disjonctif complet (on parle aussi de TDC). Dans le tableau disjonctif complet les lignes sont les individus mais les colonnes sont cette fois les modalités des variables qualitatives. Ainsi si la colonne *j* du tableau initial a K_j modalités, il y aura K_j colonnes dans le TDC, chacune correspondant à une modalité de la variable *j*. A l'intersection de la ligne *i* et de la colonne *k* on trouve y_{ik} qui vaut 1 si l'individu *i* possède la modalité k de la variable *j* et 0 sinon. Prenons un petit exemple. La variable statut matrimonial a K_j modalités et l'individu i' présente la modalité marié (qui est la 2ème modalité). Dans le tableau disjonctif complet, au niveau de la ligne *i'* et de la variable *j* on va trouver un 1 en deuxième position puisque c'est la deuxième modalité qui est possédée par l'individu *i'*, et des 0 partout

ailleurs. On retrouve ici la raison de la dénomination tableau disjonctif complet : disjonctif parce que dans chaque petit bloc on ne veut pas rencontrer deux fois un 1 et complet parce que l'on a forcément un 1. Petit point de vocabulaire une colonne est une fonction indicatrice on dit aussi plus rapidement une indicatrice. Le tableau disjonctif complet joue un rôle clé en ACM car c'est ce tableau qui est analysé par les logiciels. Cependant l'utilisateur ne le construit jamais explicitement, il fournit au logiciel le tableau initial individu x variables qualitatives.

Pour bien comprendre la structure du tableau disjonctif complet nous allons calculer ses marges. Commençons par la marge colonne. Calculons la somme des termes de la ligne *i*; rappelons que y_{ik} vaut 1 si l'individu possède la modalité *k* et 0 sinon. Donc pour chaque bloc de colonnes correspondant à une même variable, on trouve une fois et une seule la valeur 1. Ainsi en sommant sur l'ensemble des colonnes on tombe sur le nombre de variables *J*. La marge colonne est donc constante pour tous les individus et la somme des termes du tableau vaut $I \times J$, soit le nombre d'individus multiplié par le nombre de variables.

Calculons maintenant la marge ligne. Comme y_{ik} vaut 1 si l'individu possède la modalité *k*, la somme des termes de la colonne *k* est égale au nombre d'individus possédant la modalité *k*. Ce que l'on va noter de la façon suivante *I* fois p_k, c'est-à-dire le nombre d'individus multiplié par la proportion d'individus possédant la modalité *k*. Si on effectue la somme des K_j termes de cette marge ligne correspondant à une variable *j*, on trouve *I*. Et donc on retrouve que la somme des termes de l'ensemble du tableau vaut $I \times J$.

Diapositive 5

Décrivons maintenant quels sont les objectifs que l'on poursuit lorsque l'on met en œuvre une analyse des correspondances multiples. Plaçons-nous dans le cas d'une enquête. Nos individus sont des enquêtés, les variables sont les questions d'un questionnaire. On va commencer par l'étude des individus puis continuer par l'étude des variables. L'étude des individus en analyse des correspondances multiples, comme en ACP, c'est d'abord une approche multidimensionnelle. Un individu, qui est une ligne du tableau, est considéré du point de vue de l'ensemble de ses modalités. Ceci intervient lorsque l'on compare des individus : on va dire que des individus se ressemblent si dans l'ensemble ils ont le même profil. On peut avoir en tête ici un profil de type Facebook par exemple. Deux individus ont des profils similaires, et donc se ressemblent, s'ils ont choisi les mêmes modalités et deux individus sont différents s'il y a peu de modalités en commun dans leurs réponses. Alors l'ensemble de ces ressemblances et de ces différences entre individus constituent ce qu'on appelle la variabilité des individus. En analyse des correspondances multiples c'est bien la variabilité de ces individus que l'on étudie. Si tous les individus avaient répondu de la même façon au questionnaire il n'y aurait pas d'analyse statistique, il suffirait de présenter le questionnaire type. Mais justement tous les individus n'ont pas répondu la même chose et l'analyse des correspondances multiples va explorer cette variabilité, et l'explorer d'un point de vue multidimensionnel. Comme nous sommes en analyse factorielle la façon de décrire cette variabilité est d'en extraire les principales dimensions. On va ainsi mettre en évidence des dimensions qui séparent par exemple des individus extrêmes et des individus moyens. Ces dimensions seront décrites en relation avec les modalités. On ne va pas dire qu'une dimension sépare les individus 1, 3 et 4 de 10, 11 et 12, on va dire qu'on a une dimension qui sépare, par exemple, les hommes et les femmes les habitants du nord, les habitants du sud, les ouvriers, les cadres, etc., etc.

Deuxième aspect de la problématique c'est l'étude des variables. Ici nos variables sont qualitatives c'est-à-dire que l'on s'intéresse aux associations entre modalités puisque deux variables qualitatives sont liées si les modalités de l'une s'associent de façon particulière aux modalités de l'autre. On souhaite une visualisation

d'ensemble des associations entre modalités ce qui permettra d'obtenir une visualisation d'ensemble des liaisons entre les variables. Dernier aspect, la construction de variables synthétiques. On va chercher un indicateur quantitatif fondé sur les variables qualitatives qui résument le mieux possible les variables.

Toute cette problématique ressemble beaucoup à la problématique de l'analyse en composantes principales; cela tient à la structure du tableau qui est individus x variables dans les deux cas. Mais bien sûr d'un point de vue technique tout va se passer très différemment puisque dans un cas, en analyse en composantes principales, les variables sont quantitatives et dans l'autre, en analyse des correspondances multiples, les variables sont qualitatives.

Diapositive 6

Nous allons illustrer l'analyse des correspondances multiples sur les résultats d'une enquête effectuée en 2003 auprès de 8403 personnes âgées de 18 ans ou plus. L'Insee a réalisé une enquête sur la construction des identités, appelée « Histoire de vie ». Un extrait de cette base de données a été utilisé pour décrire les loisirs des français. Le jeu de données contient 8403 lignes et 22 colonnes. Les variables peuvent se répartir en deux catégories. Les 18 premières variables concernent différentes activités de loisirs. Les questions étaient posées de la façon suivante : Etes-vous allé au cinéma lors des 12 derniers mois, sans avoir été obligé de le faire ? C'est ce que nous noterons cinéma (oui/non). Nous avons également des variables Lecture (Oui / Non), Ecouter de la musique (Oui / Non), etc. Nous avons également le nombre d'heures en moyenne passées à regarder la télévision avec 5 modalités (pas du tout, 1h, environ 2h, environ 3h et 4h ou plus). Les 4 variables suivantes concernent le signalétique des individus à savoir le Sexe (homme, femme), l'âge, découpé en tranche d'âge (18-20 ans, 21-30 ans, 31-40 ans, etc.) et la situation matrimoniale (célibataire, marié, veuf, divorcé, remarié) et la catégorie socioprofessionnelle de l'enquêté (manœuvre, ouvrier, technicien, agent de maîtrise, cadre, employé, ou autre). Le tableau de données a la structure suivante : en ligne les 8403 enquêtés et en colonne les 18 activités puis les 4 variables du signalétique. La première chose que l'on regarde dans une enquête comme celle-ci ce sont les effectifs des modalités.

Diapositive 7

Ici nous donnons le nombre de pratiquants pour chaque activité, trié par ordre décroissant. Ecouter de la musique est l'activité la plus pratiquée, par 5947 personnes (sur 8403), suivie des activités lecture, marche, etc. En bas de la liste, on va trouver les activités collection et pêche qui sont moins pratiquées, et enfin on trouve le nombre de personnes pour chaque modalité du nombre d'heures passées devant la télévision (cette variable étant découpée en 5 classes). Si maintenant on regarde le signalétique, on voit que l'on a des non réponses pour la profession (1498 non réponses). Nous avons traité les non-réponses comme des modalités ordinaires. Ainsi la profession a 8 modalités.

Diapositive 8

Avant de mettre en œuvre une analyse des correspondances multiples, il convient de se demander quelle analyse précisément nous devons réaliser sachant que les variables sont constituées par deux groupes : le groupe des activités, le groupe du signalétique.

Un premier point de vue consiste à donner le statut actif aux activités et le statut supplémentaire au signalétique. Dans cette analyse, un individu est considéré uniquement du point de vue de son profil d'activité. En d'autres termes un individu c'est un profil d'activité. Tous ces profils d'activité sont le lieu d'une certaine variabilité. Et on va rechercher les principales dimensions de variabilité de ces profils d'activité. Une

fois celles-ci obtenues, on introduit le signalétique en illustratif et la question est alors de rechercher les liaisons entre les dimensions de variabilité des profils d'activités d'une part et le signalétique d'autre part.

Un deuxième point de vue symétrique du précédent consiste à donner le statut actif au signalétique et à introduire les activités en supplémentaire. Un individu est alors caractérisé par son signalétique. On est en face d'un ensemble signalétique dont on va rechercher les principales dimensions de variabilité. Cela revient à étudier finalement le plan de sondage. Puis à rechercher les liaisons entre les dimensions de variabilité du signalétique et les loisirs.

Enfin un troisième point de vue consiste à introduire les activités et le signalétique en actif, c'est-à-dire qu'un individu est composé de données hétérogènes. Sans entrer dans les détails, l'analyse d'un tel ensemble nécessite un équilibre entre les deux types de données. Ceci renvoie à d'autres méthodes comme l'analyse factorielle multiple. Nous ne dirons rien de plus sur ce point et dans la suite, nous ne présenterons que la première analyse. Un individu est donc considéré uniquement du point de vue de son profil d'activité.

Diapositive 9

L'analyse des correspondances multiples travaille à partir du tableau disjonctif complet mais ce TDC est au préalable transformé. Indiquons d'abord qu'en ACM, tous les individus ont le même poids; il n'y a aucune raison d'accorder plus d'importance aux réponses d'un enquêté plutôt qu'à celles d'un autre. Et comme la somme des poids doit être égale à 1, le poids de chaque individu est égal à $1/I$.

Rappelons aussi que la valeur y_{ik} du TDC vaut 1 si l'individu i possède la modalité k de la variable j. Cette valeur de 1 ne dépend pas de la modalité k, et en particulier elle ne dépend pas de son effectif. Or si un individu possède une modalité rare, cela le caractérise beaucoup plus que s'il possède une modalité fréquente. A la limite si une modalité est possédée par tous les individus, elle ne caractérise absolument pas l'un d'entre eux.

D'où l'idée de diviser y_{ik} par p_k. On va obtenir ainsi une valeur x_{ik} qui sera d'autant plus grande que la modalité possédée est rare.

En codant ainsi le tableau disjonctif complet, la moyenne des x_{ik} est égale à 1. En analyse des correspondances multiples, les données sont centrées ce qui veut dire que finalement, dans la case x_{ik} on va mettre $(y_{ik} / p_k) - 1$.

Le tableau des x_{ik} est le tableau sur lequel l'analyse factorielle est effectuée. Si nous faisons un parallèle avec l'analyse en composantes principales, x_{ik} est tout simplement la donnée centrée réduite. Nous avons vu quelles étaient les données et les problématiques associées de l'ACM, nous verrons dans les vidéos suivantes comment construire et ajuster un nuage d'individus et un nuage des modalités à partir de ce tableau de x_{ik}.

Deuxième partie. Visualisation du nuage des individus

(Diapositives 10 à 23)

Diapositive 10 (plan)

Nous avons vu sur quelles données s'applique l'ACM et quelles sont les problématiques associées. En ACM, comme dans toute analyse factorielle, on construit deux nuages de points : le nuage des lignes et le nuage des colonnes. Commençons ici par le nuage des lignes, c'est-à-dire le nuage des individus.

Diapositive 11

Dans le tableau disjonctif complet, un individu est représenté par une ligne du tableau. C'est un ensemble de K valeurs numériques et donc un point dans un espace à K dimensions, chaque dimension correspondant à une modalité. Chaque modalité en ACM a un poids proportionnel à son effectif. Comme la somme des poids doit être égale à 1, le poids de la modalité k est donc p_k /J. Le point M_i a comme coordonnée x_{ik}, et on a vu que chaque individu est affecté du poids 1/I.

Lorsque l'on considère l'ensemble des points on considère le nuage N_I. Ce nuage N_I a pour centre de gravité G_I qui est confondu avec l'origine puisque les variables sont centrées.

Faisons apparaître le point i' afin d'exprimer la distance entre les individus i et i'.

Le carré de cette distance s'écrit ainsi : c'est la somme des carrés des différences des coordonnées sachant que chaque carré de différence de coordonnées est pondéré par le poids de la modalité correspondante c'est-à-dire p_k /J. Si l'on exprime cette distance en fonction du tableau disjonctif complet, on obtient la relation suivante qui, après simplification, s'écrit ainsi. La distance s'écrit sous cette forme lorsque l'on présente l'analyse des correspondances multiples comme une analyse des correspondances sur le tableau disjonctif complet.

Faisons quelques remarques sur le choix de la distance. Si deux individus prennent les mêmes modalités et ont exactement le même profil, la distance qui les sépare est nulle. Si deux individus ont en commun beaucoup de modalités, la distance qui les sépare sera petite. Si deux individus ont en commun beaucoup de modalités, sauf une, rare, qui est prise par l'un des deux, alors la distance qui les sépare sera relativement grande grâce au p_k qui sera petit pour prendre en compte la spécificité de l'un des deux. Si deux individus ont en commun une modalité rare, la distance qui les sépare sera relativement petite pour prendre en compte leur spécificité commune.

Calculons maintenant la distance d'un point à l'origine. Le carré de la distance entre un individu et l'origine est égal à la somme des carrés des coordonnées toujours pondérées par le poids des modalités. Si l'on fait apparaître le tableau disjonctif complet, on a la relation suivante qui après simplification s'écrit ainsi. On voit bien apparaître ici une somme des y_{ik} / p_k, c'est-à-dire une somme qui va être d'autant plus grande que les modalités possédées par l'individu i sont rares c'est-à-dire associées à des p_k qui sont petits. Cette relation met en évidence qu'un individu est d'autant plus loin de l'origine qu'il possède des modalités rares.

Calculons pour terminer l'inertie totale du nuage N_I. On calcule d'abord l'inertie d'un point, c'est-à-dire son poids multiplié par le carré de sa distance à l'origine, et pour le nuage on effectue la sommation sur tous les individus. Tout calcul fait, on obtient que l'inertie totale est égale à K / J - 1. Cette quantité ne dépend pas du

contenu du tableau lui-même mais simplement d'un aspect de son format c'est-à-dire le nombre de modalités et le nombre de variables. Ce résultat est différent de celui de l'analyse des correspondances puisqu'en analyse des correspondances, l'inertie totale est égale au phi2 et mesure donc l'écart à l'indépendance. Par contre ce résultat est assez analogue à celui de l'analyse en composantes principales normée où l'inertie totale est égale au nombre de variables et ne dépend donc pas du contenu du tableau mais uniquement de son format.

Diapositive 12

Nous avons défini le nuage des individus. Comme pour toute analyse factorielle, nous allons maintenant projeter ce nuage pour le visualiser dans un sous-espace de dimension plus petite, souvent on se contentera de représenter le nuage sur un plan. Pour ce faire, nous appliquons une analyse factorielle à ce nuage, c'est-à-dire que l'on projette ce nuage sur une suite d'axes orthogonaux d'inertie maximum. On obtient un plan factoriel pour les individus qui est la meilleure représentation plane des individus.

Diapositive 13

Reprenons le jeu de données sur les loisirs des Français. Les 18 premières variables correspondent aux loisirs des Français et sont considérées comme les variables actives du jeu de données. Ainsi, ce sont ces 18 variables qui servent à calculer les distances entre individus. Les variables de signalétique ont le statut de variables supplémentaires et seront utilisées ultérieurement pour aider à l'interprétation des résultats.

Diapositive 14

Comme dans toute analyse factorielle, commençons par regarder la décroissance des inerties associées à chaque axe. Ici le diagramme en barres met en évidence que le premier axe a une inertie supérieure aux suivantes. Cela incite à interpréter le premier plan factoriel en priorité et à s'y limiter dans un premier temps. Le pourcentage d'inertie associé à ce premier plan est de 24%. Ce chiffre peut-il être considéré comme élevé ou bas? Rappelons que nous analysons des profils d'activités assez complexes puisqu'il y a 18 loisirs et une très grande variabilité. Parmi nos enquêtés il y a des jeunes, des vieux, des hommes, des femmes, on peut s'attendre à une très grande diversité des profils d'activités. Et on ne peut pas s'attendre à ce que cette diversité puisse s'exprimer simplement sur deux composantes. Si l'on regarde maintenant les axes 3 et 4, ils ont des inerties comparables et relativement grandes. Dans cette analyse, nous ne regarderons pas le plan 3-4 mais il s'interprète bien. Finalement si l'on interprétait les 4 axes, on aurait un pourcentage d'inertie expliqué de 35%. Le reste, c'est un ensemble de variabilité individuelle.

Diapositive 15

En analyse des correspondances multiples, comme en analyse en composantes principales, la première chose qu'il faut examiner est l'allure générale du nuage des individus. Ici cette allure générale est tout à fait régulière. Il n'y a rien d'autre à dire. Alors qu'est-ce que ça serait une allure très particulière?

Diapositive 16

Cette représentation par exemple est particulière et met en évidence trois classes d'individus bien distinctes. Lorsqu'on a vu ça il est clair que ces trois classes d'individus vont dominer l'interprétation. C'est pour ça qu'il faut toujours examiner en premier l'allure générale du nuage des individus, quitte à dire qu'elle est tout à fait régulière et qu'il n'y a rien de particulier d'autre à en dire.

Cette autre représentation des individus montre une allure de graphe particulière, mais assez fréquente en ACM, en forme de fer à cheval. On parle d'effet Guttman. Cet effet Guttman a tendance à répartir sur un axe, ici l'axe horizontal, les individus selon des modalités croissantes (par exemple faibles à gauche, moyennes au milieu, fortes à droite) et à opposer sur l'autre axe, ici l'axe vertical, les individus extrêmes (qui prennent de faibles ou fortes valeurs) aux individus moyens.

Diapositive 17

Pour comprendre les principales dimensions de variabilité, nous pouvons utiliser les variables qualitatives du jeu de données. Une première idée consiste à colorier les points en fonction des modalités qu'ils prennent sur une variable. Par exemple, ici pour la variable jardinage, ceux qui jardinent sont en rouge, les autres en noir. On voit que les points rouges sont plutôt en haut du graphe par rapport aux points noirs, et donc l'axe 2 semble séparer les individus selon qu'ils jardinent ou non. Evidemment, on peut construire ces graphes pour chaque variable mais cela deviendrait vite fastidieux.

On peut alors avoir l'idée d'utiliser les modalités des variables, ou les variables, pour caractériser et interpréter le graphe des individus. Commençons par utiliser les modalités.

Une modalité peut être vue comme un groupe d'individus et il est donc naturel de la représenter directement sur le graphe des individus. Pour ce faire, on va placer une modalité au barycentre des individus qui la possèdent. Pour la variable jardinage, on a donc 2 points : la modalité "jardinage oui" en rouge et "jardinage non" en noir. Rappelons la propriété suivante : l'origine du nuage est au barycentre des modalités d'une même variable si chaque modalité a un poids proportionnel à son effectif. Ici, le point jardinage "non" est plus proche de l'origine que le point jardinage "oui". Cela signifie qu'il y a peu de personnes qui font du jardinage comparé au nombre de personnes qui n'en font pas : 3356 disent jardiner et 5047 ne pas jardiner et ces effectifs sont proportionnels à la distance entre la modalité et l'origine. Ainsi les modalités de faible effectif seront plus éloignées de l'origine.

Diapositive 18

Dans ce graphe, on a placé toutes les modalités de toutes les variables sur la représentation des individus. Comme les modalités sont des moyennes (puisque ce sont des barycentres), les points sont tous proches du centre de gravité. Les modalités "oui" sont coloriées en rouge et les modalités "non" en noir.

Diapositive 19

Si on efface les points représentant les individus et que l'on zoome sur le graphique, voici la répartition des modalités sur le premier plan factoriel du graphe des individus. La première chose qui frappe dans ce graphique c'est bien sûr le regroupement d'un côté des modalités activités "oui" et de l'autre côté des modalités activités "non". Indiquons tout de suite qu'il ne s'agit pas d'un artefact car, dans le tableau disjonctif complet, rien n'indique que certaines modalités sont des "oui" et d'autres des "non". Il s'agit donc bien d'une structure dans les données. Cette opposition s'organise non pas selon le premier axe ou le deuxième axe mais plutôt selon la première bissectrice. Il n'y a pas d'inconvénients à interpréter de préférence cette première bissectrice.

Diapositive 20

Pour nous aider dans l'interprétation, nous allons prendre quelques individus caractéristiques. Choisissons 4 individus qui sont bien opposés par cette bissectrice : 5938 et 2432 d'un côté et 8325 et 203 de l'autre. Le graphique des modalités montre que les individus 5938 et 2432 sont du côté des réponses "oui", ce qui suggère qu'ils ont beaucoup d'activités de loisirs. A contrario, les individus 8325 et 203 sont du côté des modalités "non", ce qui suggère qu'ils ont peu d'activités. Si on revient au tableau de données, on retrouve parfaitement cette information : 5938 et 2432 font presque toutes les activités tandis que 8325 et 203 n'en font aucune (mise à part regarder beaucoup la TV). Cette opposition sur la première bissectrice correspond donc à une intensité d'activités de loisirs et oppose des individus qui ont déclaré avoir beaucoup de loisirs à des individus qui ont déclaré en avoir très peu.

Diapositive 21

Après avoir interprété la première bissectrice de ce plan factoriel, nous allons essayer d'interpréter la seconde bissectrice. Cette seconde bissectrice sépare deux groupes de modalités. Un premier groupe comprend les modalités cinéma "oui", ordinateur "oui", spectacle "oui", sport "oui", jouer de la musique "oui", jardinage "non", cuisine "non", tricot "non". Ce groupe de modalités s'oppose à un autre groupe qui comprend les modalités inverses cinéma "non", ordinateur "non", spectacle "non", sport "non", jouer de la musique "non", jardinage "oui", cuisine "oui", tricot "oui". Utilisons à nouveau des individus particuliers pour caractériser cette seconde bissectrice. Les individus 255 et 6766 ont des loisirs plutôt jeunes et n'ont pas de loisirs disons plutôt tranquilles. A l'opposé, 1143 et 5676 ont plutôt des loisirs tranquilles et pas d'activités jeunes. On peut noter que les 4 individus pratiquent à peu près autant d'activités et c'est pour cette raison que sur la première bissectrice ils ont à peu près la même coordonnée. On est bien ici en train de commenter une bissectrice orthogonale à la première et donc de répondre à la question : à nombre d'activités constant, qu'est-ce qui sépare nos individus ? Si on fait un bilan de cette représentation des modalités, on peut dire que la variabilité des profils d'activités peut être organisée autour de deux dimensions : une dimension d'intensité qui oppose les individus qui ont beaucoup d'activités à des individus qui en ont peu et une dimension sur les types d'activités (jeunes ou seniors). Nous verrons ultérieurement comment projeter les modalités des variables supplémentaires (du signalétique) afin de répondre à la question suivante : est-ce que ces dimensions de variabilité que nous venons de mettre en évidence sont liées à certaines variables du signalétique, et notamment à l'âge par exemple.

Diapositive 22

Reprenons le graphe des individus coloriés en fonction des modalités de la variable jardinage. On va considérer les projetés de chaque individu sur un axe.

On voit ici tous les projetés sur l'axe 2. Et on va calculer un indicateur de liaison entre les coordonnées des individus sur l'axe 2 et la variable qualitative jardinage.

L'indicateur de liaison entre une variable quantitative et une variable qualitative est le rapport de corrélation. C'est cet indicateur qui est utilisé en analyse de variance à 1 facteur. On utilise souvent le carré du rapport de corrélation qui est égal au pourcentage de variabilité de la variable quantitative expliquée par la variable qualitative. Cet indicateur varie entre 0 et 1 : il vaut 0 s'il n'y a aucune liaison entre les 2 variables, i.e. si les moyennes de la variable quantitative pour chaque modalité de la variable qualitative sont égales. Et il vaut 1 si les moyennes sont différentes et si la variabilité intra-modalité est nulle, autrement dit si tous les individus d'une même modalité prennent exactement la même valeur sur la variable quantitative.

Ici le rapport de corrélation au carré entre la variable jardinage et la seconde dimension est égal à 0.453. Ce graphe ne permet pas de lire directement le carré du rapport de corrélation mais il donne une idée de la séparation des points selon les modalités de la variable jardinage. Et un carré de rapport de corrélation de 0.453 sépare déjà bien les points.

Le carré du rapport de corrélation entre Jardinage et la première dimension est égal à 0.047. Cette valeur est faible et en effet, sur le premier axe, les projetés rouges et noirs semblent bien mélangés; il n'y a pas de séparation de ces 2 classes d'individus.

Diapositive 23

On récupère alors les rapports de corrélation au carré de chaque variable avec les dimensions 1 et 2 pour construire un graphe des variables que l'on appelle graphe du carré des liaisons. Tous les points vont en effet se retrouver dans un carré de côté 1. Notez que, sur ce graphe, une variable qui a 3 modalités ou plus peut avoir à la fois une coordonnée de 1 sur la première et sur la 2ème dimension. Pour la variable jardinage, on place le point aux coordonnées 0.047 sur l'axe 1 et 0.453 sur l'axe 2. On va alors commenter les variables qui ont les carrés de rapport de corrélation les plus grands sur les axes 1 et 2 pour démarrer l'interprétation. Mais on va aussi prendre en compte la valeur de chaque rapport de corrélation au carré car une variable, même si elle est la plus liée à un axe, ne sera pas vue de la même manière si le carré de son rapport de corrélation vaut 0.3 ou 0.9.

Dans notre exemple, les liaisons ne sont pas très fortes, pas très proches de 1. Si on zoome sur le graphe, on voit que les variables Exposition, Spectacle, Ordinateur, Cinéma sont liées à la première dimension. Le carré du rapport de corrélation varie entre 0.3 et 0.4. Ceci n'est pas très élevé mais compte-tenu du nombre très important d'individus, ces rapports de corrélation au carré sont (très) significativement différents de 0. On peut donc prendre en compte ces variables dans l'interprétation du premier axe et donc du plan. Attention toutefois on ne peut pas interpréter les probabilités critiques exactement comme dans un test classique. En effet, l'axe factoriel a été construit à partir des variables qualitatives, donc il ne faut pas s'étonner qu'il y ait des liaisons et que les carrés de rapport de corrélation soient grands. On a tout fait pour. Autrement dit cette probabilité critique ne peut s'interpréter comme un test classique de statistique que pour les variables supplémentaires. Cela reste cependant un indicateur tout à fait intéressant.

Nous pouvons noter une propriété très intéressante : l'ACM recherche des dimensions factorielles les plus liées possibles aux variables du jeu de données au sens du carré du rapport de corrélation. Concrètement, pour trouver l'axe s, on cherche parmi les variables orthogonales aux axes précédemment trouvés, la variable la plus liée aux variables du jeu de données, plus liée au sens où elle maximise la somme des rapports de corrélation au carré. Nous verrons dans la prochaine vidéo comment construire un nuage des modalités et comment en avoir une représentation optimale. Nous soulignerons aussi le lien avec la représentation des individus que nous venons d'obtenir. C'est cette liaison entre les 2 représentations qui fait toute la richesse de l'ACM.

Troisième partie. Visualisation du nuage des modalités et représentation simultanée

(Diapositives 24 à 29)

Diapositive 24 (plan)

Nous avons vu comment construire et ajuster un nuage des individus et comment l'interpréter grâce aux modalités des variables. Dans cette vidéo, nous allons voir comment construire un nuage des modalités et comment en avoir une représentation optimale. Nous verrons le lien entre la représentation optimale des individus et la représentation optimale des modalités grâce aux relations de transition.

Diapositive 25

Reprenons le tableau qui croise les individus et les modalités. Sur la ligne *i* et la colonne *k*, on a x_{ik} qui est égal à y_{ik}/ p_k -1 (y_{ik} vaut 1 si l'individu prend la modalité k, 0 sinon; p_k est la proportion d'individus prenant la modalité *k*; et on soustrait 1 pour centrer chaque colonne). Examinons le nuage des modalités, c'est-à-dire celui des colonnes. Une colonne prend une valeur pour chacun des *I* individus, c'est donc un ensemble de *I* valeurs numériques. En ce sens c'est un point d'un espace à *I* dimensions, chaque dimension correspondant à un individu. Voici ici l'individu *i* auquel est affecté le poids 1/*I*. La modalité est représentée par le point M_k à partir de ses coordonnées x_{ik}. On rappelle que le poids d'une modalité est proportionnel à son effectif.

Lorsque l'on considère l'ensemble des modalités on considère le nuage N_K des modalités. Ce nuage possède la propriété suivante : pour un individu *i*, la somme des coordonnées x_{ik} des modalités d'une même variable est égale à 0. Ainsi l'origine est confondue avec le centre de gravité des modalités d'une même variable. Comme cette propriété est vraie pour toutes les variables, le centre de gravité du nuage N_K est confondu avec l'origine.

Calculons la variance de la modalité k. Cette variance est égale au carré de la distance entre cette modalité et l'origine dans l'espace R^I. Ecrivons cette distance au carré : c'est bien la somme des carrés des coordonnées pondérées par les poids 1/*I*. Exprimons cette variance en fonction du tableau disjonctif complet. Voici ce que l'on obtient. Tout calcul fait on obtient $1/ p_k$ -1. C'est-à-dire que la distance entre un point M_k représentant donc la modalité k et l'origine est d'autant plus grande que cette modalité est rare. En analyse des correspondances multiples, les modalités rares sont très éloignées de l'origine.

Regardons comment varie cette distance en fonction de pk pour quelques valeurs numériques. On voit que lorsque pk passe de 1/2 à 1/5, la distance double; lorsque pk passe de 1/5 à 1/10, la distance augmente encore de 50%. C'est donc un point important et l'on voit que cette distance augmente beaucoup avec la rareté d'une modalité. Mais attention, en analyse factorielle c'est l'inertie, et non la distance, qui compte dans la construction des axes.

Examinons l'inertie de la modalité *k*. L'inertie est le carré de la distance entre *k* et l'origine multiplié par le poids de la modalité. On voit bien que lorsqu'une modalité est rare la distance augmente mais le poids diminue. Il y a donc un antagonisme. Tout calcul fait, on obtient: $(1- p_k)/J$, ce qui montre bien que lorsqu'une modalité est rare, elle a une forte inertie. Elle va donc influencer les résultats de l'analyse des correspondances multiples.

Regardons avec J=10 variables comment évolue cette inertie en fonction de la fréquence. Lorsque pk passe de 1/2 à 1/10, l'inertie double pratiquement puisqu'elle passe de 0.05 à 0.09. On voit bien ici l'influence d'une modalité rare. En revanche remarquons que lorsque l'on passe de 1/10 à 1/100, c'est-à-dire d'une modalité rare à une modalité très rare, l'inertie n'augmente quasiment plus. Le problème en ACM est donc bien celui des modalités rares qui ont une forte influence plus que celui des modalités très rares.

Calculons enfin la distance entre deux modalités et exprimons cette distance en fonction des éléments du tableau disjonctif complet y_{ik} et $y_{ik'}$. On peut exprimer cette distance uniquement en fonction de $p_k * p_{k'}$ et de la proportion d'individus qui prennent à la fois les modalités *k* et *k'*, ce que nous notons $p_{kk'}$. On voit que plus on a d'individus qui ont choisi une seule des 2 modalités et plus la distance est grande.

Diapositive 26

Calculons maintenant l'inertie d'une variable *j*. Il suffit en fait de faire la somme des modalités de la variable. On reprend donc l'inertie d'une modalité *k* et on somme sur les K_j modalités de la variable *j*. Tout calcul fait, on trouve $(K_j - 1)/J$, c'est-à-dire que l'inertie d'une variable est proportionnelle au nombre de modalités moins 1 de la variable.

Dans un questionnaire par exemple, la variable sexe a 2 modalités et donc une inertie de 1/*J* (1 sur le nombre de variables). Prenons maintenant la variable région, qui a 21 modalités, l'inertie totale de cette variable région vaut 20/*J*. Dans un premier temps on peut s'inquiéter de cette disparité entre ces deux inerties qui varient de 1 à 20 et chercher à avoir des variables ayant à peu près le même nombre de modalités.

En fait, il n'en est rien car l'analyse des correspondances multiples gère parfaitement cette différence d'inertie. En effet, une variable j à K_j modalités correspond à K_j indicatrices dans le tableau disjonctif complet et correspond donc à un sous-espace à K_j -1 dimensions. On peut voir qu'il y a K_j -1 dimensions en considérant qu'il y a une liaison entre toutes ces indicatrices (leur somme vaut 1).

Or nous avons vu que l'inertie de l'ensemble des modalités de la variable j est proportionnelle à K_j - 1, K_j - 1 étant précisément la dimension du sous-espace dans lequel se trouve cette inertie. Autrement dit une variable qui a beaucoup de modalités a une inertie très importante mais cette inertie est en quelque sorte diluée dans un sous-espace de dimension également importante. Ceci explique pourquoi la variable sexe ne peut être très liée qu'à un seul axe. Il n'y aura qu'un seul axe qui oppose fortement les hommes et les femmes. Alors que la variable région, elle, pourra être très liée à de nombreux axes, un axe opposant le nord et le sud, un autre l'est à l'ouest, etc., etc.

Enfin, on peut calculer l'inertie totale en faisant la somme des inerties de chaque variable. Et on retombe sur la même inertie que l'inertie du nuage des individus, à savoir K/J - 1.

Diapositive 27

On va, comme on en a pris l'habitude, ajuster ce nuage des modalités par analyse factorielle, c'est-à-dire construire séquentiellement les axes en recherchant un axe qui maximise l'inertie et qui est orthogonal aux axes précédemment trouvés.

Voici la représentation des modalités. C'est cette représentation qui est optimale pour visualiser le nuage des modalités. Ce graphe ressemble très fortement à celui que nous avons construit lorsque nous avons

utilisé les modalités pour interpréter le graphe des individus même si la représentation n'est pas exactement la même. Cependant, l'interprétation de ce graphe sera très similaire à celle que nous avons déjà faite : un premier axe, ou plutôt une première bissectrice, qui oppose ceux qui ont peu d'activités de ceux qui en ont beaucoup. Et, une deuxième bissectrice qui oppose des activités plutôt jeunes à des activités plutôt seniors.

Diapositive 28

De façon symétrique par rapport à ce que nous avions fait dans l'étude des individus, nous pouvons projeter les individus sur ce graphe des modalités. Un individu peut être considéré comme un ensemble de modalités, celles qu'il possède, et on peut positionner un individu au barycentre des modalités qu'il possède. Tous les individus se positionnent au centre du nuage, ce qui est attendu puisque les individus sont ici considérés comme des moyennes.

Diapositive 29

Revenons sur les différents graphes que nous avons construits et voyons le lien entre tous ces graphes. Reprenons tout d'abord la représentation optimale du nuage des individus que nous avons construite dans l'étude des individus (dans la vidéo précédente).

Pour interpréter ce graphe, nous avons positionné chaque modalité au barycentre des individus qui la possède. Si on note $G_s(k)$ la coordonnée de la modalité *k* sur la dimension *s*, nous pouvons écrire cette formule. y_{ik} vaut 1 si l'individu i prend la modalité *k* et 0 sinon, donc on a ici la somme des coordonnées des individus qui prennent la modalité *k*, divisée par le nombre d'individus qui prennent la modalité *k*. On est bien en train de calculer la moyenne des coordonnées des individus qui prennent la modalité *k*.

Nous pouvons maintenant considérer la représentation optimale des modalités que nous avons construite dans cette vidéo. Nous pouvons remarquer que les coordonnées des modalités sont dilatées par rapport au graphe de gauche.

Nous pouvons projeter chaque individu au barycentre des modalités qu'il possède. La coordonnée d'un individu i sur l'axe s est donc la moyenne des modalités prises par l'individu. Dans cette formule, y_{ik} que multiplie $G_s(k)$ correspond à la somme des coordonnées des modalités prises par l'individu *i*, et comme un individu possède une et une seule modalité par variable et qu'il y a *J* variables, on divise par *J* pour avoir la moyenne des coordonnées des modalités que l'individu possède. Les 2 graphes de droite et de gauche diffèrent, mais ...

... à une dilatation près. Dans le graphe de gauche, si on conserve la position des individus mais que l'on dilate le nuage des modalités d'un coefficient 1 sur racine de λ_s , voici ce que l'on obtient comme graphe. λ_s est la valeur propre associée à l'axe s et je dis que l'on dilate le nuage des modalités car λ_s est toujours inférieure à 1. Dans l'écriture de la relation, il suffit d'introduire ce coefficient 1 sur racine de λ_s .

Sur le graphe de droite, nous conservons la position des modalités et dilatons cette fois le nuage des individus, toujours en utilisant le même coefficient 1 sur racine de λ_s . Dans la relation de transition, on introduit donc seulement un terme 1 sur racine de λ_s . Et que remarque-t-on maintenant ? Avec ces 2 dilatations, les graphes de gauche et de droite sont parfaitement identiques.

C'est ce graphe, que l'on appelle représentation simultanée et qui est fourni par les logiciels quand on exécute une ACM. La représentation des individus est optimale, celle des modalités également. Les deux relations de transition montrent comment passer d'une représentation à l'autre et nous aide dans l'interprétation du graphe. La relation de gauche nous dit qu'une modalité est au pseudo-barycentre des individus qui la possèdent et celle de droite nous dit qu'un individu est au pseudo-barycentre des modalités qu'il possède. Ainsi, un individu sera du côté des modalités qu'il possède et à l'opposé des modalités qu'il ne possède pas. De même une modalité est du côté des individus qui la possède et à l'opposé de ceux qui ne la possède pas. Cette symétrie entre les deux propriétés fait que l'on parle de double propriété barycentrique. Nous avons vu comment ajuster le nuage des individus et le nuage des modalités, nous verrons dans la prochaine vidéo plusieurs aides à l'interprétation et comment utiliser des informations supplémentaires comme le signalétique en ACM.

Quatrième partie. Aides à l'interprétation

(Diapositives 30 à 39)

Diapositive 30 (plan)

Nous avons vu comment ajuster le nuage des individus et le nuage des modalités, dans cette vidéo nous allons détailler quelques aides à l'interprétation commune à toute analyse factorielle, voir comment utiliser des informations supplémentaires comme le signalétique en ACM.

Diapositive 31

L'inertie d'un axe en ACM est particulière car elle est égale à la moyenne des carrés des rapports de corrélation entre l'axe et les variables. Cette propriété valide l'interprétation des facteurs de l'ACM en tant que variable synthétique. Les facteurs de l'ACM sont des variables quantitatives qui synthétisent les variables qualitatives. On peut aussi noter que, comme les carrés de rapport de corrélation sont compris entre 0 et 1, la contribution d'une variable à un axe est bornée par 1, ce qui montre en quel sens les variables sont équilibrées en ACM. De plus, la moyenne des rapports de corrélation, et donc l'inertie, est aussi toujours comprise entre 0 et 1.

Si on s'intéresse aux pourcentages d'inertie maintenant. Ils sont généralement plus faibles qu'en ACP ou en AFC car les individus évoluent dans un espace de dimension élevée *K-J*, d'autant plus élevée que le nombre de modalités par variable est grand. Le petit calcul suivant montre pourquoi le pourcentage d'inertie associée à une dimension est souvent faible. On calcule l'inertie d'un axe, λ_s, sur l'inertie totale, sur $(K-J)/J$; et comme la valeur propre λ_s est inférieure à 1, alors le pourcentage d'inertie est inférieur à $J/(K-J)\times 100$. Avec 10 variables et 10 modalités par variable, même si les variables sont extrêmement liées, le pourcentage d'inertie maximum expliqué par un axe est de 11%. Un autre petit calcul montre que l'inertie moyenne est égale à 1 sur le nombre de variables. Cette valeur peut aider à décider combien de dimensions interpréter en ACM. On évitera d'interpréter des dimensions qui ont une inertie inférieure à 1/*J*.

Diapositive 32

Les aides à l'interprétation contribution et qualité de représentation se calculent comme pour les autres méthodes d'analyse factorielle, ACP ou AFC. Cependant, les qualités de projection sont généralement faibles sur chaque axe, ce qui est attendu puisqu'il y a beaucoup de dimensions. Quant aux modalités, comme un poids est associé à chaque modalité, les contributions importantes ne correspondent pas nécessairement aux points les plus éloignés sur le graphe.

La contribution d'une variable à la construction d'un axe se calcule en sommant les contributions de toutes ses modalités. Cette contribution est égale au rapport de corrélation au carré entre l'axe et la variable divisé par le nombre de variables. Pour calculer la contribution relative de la variable, on divisera par l'inertie de l'axe λ_s.

Diapositive 33

Nous pouvons maintenant voir comment utiliser des informations supplémentaires pour interpréter le graphe de la représentation simultanée. Pour ce faire, nous utilisons les relations de transition pour les calculs des coordonnées des éléments supplémentaires. Par élément, nous entendons soit une ligne soit une colonne d'un tableau c'est-à-dire ici soit un individu soit une modalité. Dans toute analyse factorielle, un élément est dit supplémentaire s'il n'est pas utilisé dans la construction des axes. Précisons à propos des modalités que toutes les modalités d'une même variable doivent avoir le même statut : elles doivent être toutes actives ou toutes supplémentaires. C'est pourquoi on parle de variables actives et de variables supplémentaires. Les propriétés barycentriques servent à calculer la position d'un individu ou d'une modalité supplémentaire. Prenons le cas d'une modalité supplémentaire. Une fois la position des individus fixée, il est possible de calculer le barycentre d'un sous-ensemble quelconque d'individus. En pratique les ACM comportent presque toujours des variables supplémentaires car ces variables donnent des éléments du contexte de l'analyse et sont donc excessivement précieuses.

Le graphique suivant donne les modalités des variables supplémentaires. Afin de ne pas surcharger le graphique, nous n'avons pas reproduit les modalités actives et les individus. Cependant les individus actifs ou supplémentaires, les modalités actives ou supplémentaires peuvent être positionnés sur le même graphe. On voit que les variables de signalétique sont liées aux dimensions factorielles.

Les classes d'âge, par exemple, s'organisent dans un ordre croissant en partant du bas à droite, en remontant et en terminant sur la gauche. De même, les professions ou les statuts matrimoniaux s'organisent le long des dimensions factorielles. On va donc interpréter le graphique d'ACM en disant qu'en bas à droite on trouve des individus jeunes, plutôt des cadres, vivant seuls, et qui ont beaucoup d'activités de type cinéma, spectacle, ordinateur. En haut, les individus sont plus âgés, et ont des activités comme la pêche, le tricot, le jardinage. Et enfin, sur la gauche, on trouve des ouvriers, manœuvres, plus âgés, qui ont peu d'activités de loisirs. Pour cette interprétation nous avons travaillé en deux temps : dans un premier temps, on travaille uniquement à partir des variables actives et dans un deuxième temps, on introduit les variables supplémentaires.

Diapositive 34

Nous situons dans le thème aides à l'interprétation la notion de variables supplémentaires quantitatives. En ACM, les variables actives sont nécessairement qualitatives donc les variables quantitatives sont nécessairement supplémentaires. Dans notre exemple, nous avions interprété la première bissectrice comme le nombre d'activités de loisir pratiquées. D'où l'idée, pour valider cette interprétation de calculer ce nombre d'activités pratiquées par individu et de créer une variable quantitative "nombre d'activités". Pour représenter cette variable on va procéder comme en ACP et utiliser la représentation dite du cercle des corrélations : on représente une variable par ces coefficients de corrélation avec les facteurs. On obtient le graphe suivant qui montre que le nombre d'activités pratiquées est parfaitement représenté sur ce premier plan. Il correspond en gros à la première bissectrice, ce qui achève de valider notre interprétation.

Si on souhaite qu'une variable quantitative soit active, il est possible de la discrétiser en classes, de la rendre ainsi qualitative et de l'utiliser comme variable active.

Diapositive 35

Comme en ACP, nous pouvons décrire les dimensions factorielles de l'ACM à partir des variables actives ou illustratives, quantitatives ou qualitatives. Pour les variables quantitatives, on calcule le coefficient de corrélation entre une variable et une dimension et on trie les coefficients de corrélation. Ici, la variable nombre d'activités est très liée à la première dimension. Pour les variables qualitatives, on va construire un modèle d'analyse de variance par variable et expliquer les coordonnées sur la dimension en fonction de la variable. On construit le test de Ficher pour voir la liaison globale entre la variable et la dimension, puis les tests de Student pour voir quelles sont les modalités qui ont des coordonnées particulières sur les dimensions. Ici, toutes les probabilités critiques sont extrêmement faibles car le nombre d'individus est important dans le jeu de données.

Diapositive 36

Il y a une autre façon de présenter les données lorsque l'on est face à un ensemble de variables qualitatives, c'est le tableau de Burt. L'idée de ce tableau est de croiser toutes les variables avec elles-mêmes. Dans le tableau que nous avons représenté ici, chaque rectangle est un tableau croisé ou tableau de contingence entre 2 variables. Les lignes et les colonnes d'un tableau de contingence sont les modalités. Cela veut dire que le tableau de Burt croise les modalités de toutes les variables avec elles-mêmes. Le tableau de Burt récapitule l'ensemble des liaisons entre les variables prises 2 à 2. Il y a une analogie ici avec la matrice des corrélations qui elle aussi rassemble les corrélations entre toutes les variables prises 2 à 2.

Quelle analyse mettre en œuvre pour analyser le tableau de Burt ? Une idée simple, par analogie à l'analyse des correspondances sur le tableau de contingence, consiste à mettre en œuvre une analyse des correspondances sur le tableau de Burt. Lorsque l'on applique un programme d'AFC au tableau de Burt, on obtient exactement les mêmes facteurs que l'ACM mais avec des valeurs propres différentes. Les valeurs propres issues du tableau de Burt sont le carré des valeurs propres issues du tableau disjonctif complet. Quelles valeurs propres faut-il choisir? Nous avons donné une interprétation des valeurs propres issues du tableau disjonctif complet en tant que moyenne des carrés des rapports de corrélation, cette interprétation est tout à fait utile et c'est celle-ci que nous choisissons de conserver.

Finalement cette analogie nous apprend une chose, c'est que l'ACM ne dépend que des liaisons entre les variables prises 2 à 2. Il y a là encore une analogie avec l'ACP qui ne dépend que de la matrice des corrélations.

Diapositive 37

Conclusion sur l'analyse des correspondances multiples. Le premier résultat c'est que l'ACM est la méthode factorielle adaptée au tableau individus x variables qualitatives. La problématique générale de l'ACM est de rechercher les principaux facteurs de variabilité des individus, ou encore une synthèse des liaisons entre les variables. A ce niveau de généralité cette problématique est exactement identique à celle de l'ACP. Sur un plan technique on peut dire que les principales règles d'interprétation sont très simples. Un individu est au barycentre de ses modalités et une modalité est au barycentre des individus qui la possèdent. On a donc une méthode qui est très générale parce que les tableaux individus x variables qualitatives sont très fréquents avec des règles d'interprétation très simples. L'analyse des correspondances multiples est une méthode particulièrement adaptée au traitement des enquêtes.

L'interprétation des valeurs propres comme moyenne des carrés de rapport de corrélation est une de propriété de l'ACM. Cette propriété valide l'interprétation des facteurs de l'ACM en tant que variable synthétique. Les facteurs de l'ACM sont des variables quantitatives qui synthétisent les variables qualitatives.

On a insisté sur le carré des liaisons, cette représentation est précieuse en particulier lorsqu'il y a beaucoup de variables.

Comme pour les autres méthodes d'analyse factorielle, il est utile de valider ses interprétations des résultats d'une ACM en revenant aux données. L'ACM va suggérer des liaisons entre variables qualitatives et il est possible de construire un tableau croisé de 2 variables et d'analyser et visualiser ce tableau par AFC.

Enfin la convergence entre l'analyse du tableau disjonctif complet et celle du tableau de Burt est un argument en faveur de l'intérêt de la méthode. C'est toujours intéressant de voir que différents points de vue conduisent au même résultat.

Enfin, l'ACM permet de passer d'un tableau de variables qualitatives à un tableau de dimensions factorielles, c'est-à-dire de variables quantitatives. Cette méthode peut donc être vue comme un prétraitement avant une classification par exemple.

Diapositive 39

Pour terminer indiquons quelques références en phase avec cette présentation de l'ACM. Vous avez vu toutes les vidéos de cours sur l'ACM, vous pouvez maintenant voir la vidéo qui décrit comment mettre en œuvre une ACM sous FactoMineR.