

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

- Première partie. Données - problématique et modèle d'indépendance**
Diapositives 1 à 14
Pages 2 à 6
- Deuxième partie. Visualisation des nuages des lignes et des colonnes**
Diapositives 15 à 21
Pages 7 à 10
- Troisième partie. Inertie et pourcentage d'inertie**
Diapositives 22 à 26
Pages 11 à 13
- Quatrième partie. La représentation simultanée**
Diapositives 27 à 32
Pages 14 à 17
- Cinquième partie. Aides à l'interprétation**
Diapositives 33 à 42
Pages 18 à 21

Première partie. Données - problématique

(Diapositives 1 à 14)

Diapositive 1

Cette semaine, nous vous présentons 5 vidéos de cours sur l'analyse factorielle des correspondances.

Diapositive 2 (plan)

L'ensemble des vidéos aborde les points suivants : nous commencerons par décrire les données, préciserons quelques notations et les questions que l'on se pose lorsque l'on met en œuvre une analyse des correspondances. Nous verrons que la notion cruciale en analyse des correspondances est celle de la liaison entre deux variables qualitatives et qu'étudier la liaison entre 2 variables qualitatives c'est examiner l'écart entre les données et une situation d'indépendance. Nous verrons ainsi comment l'analyse appréhende l'écart à l'indépendance. Nous raisonnerons de façon essentiellement géométrique en construisant un nuage des lignes et un nuage des colonnes, nuages que nous représenterons par analyse factorielle c'est-à-dire en projetant sur des plans en pratique. Nous donnerons quelques indications sur les pourcentages d'inertie. De ce point de vue l'analyse des correspondances ne se distingue pas d'autres méthodes d'analyse factorielle comme l'analyse en composantes principales. Nous indiquerons également quelques résultats techniques concernant les inerties, et là il s'agit de résultats qui sont tout à fait particulier en analyse des correspondances. De même en analyse des correspondances la représentation simultanée des lignes et des colonnes est assez particulière et elle bénéficie des propriétés dites barycentriques. Nous décrirons enfin quelques aides à l'interprétation, contribution et qualité de représentation. De ce point de vue, l'analyse des correspondances n'est pas particulière par rapport à l'analyse en composantes principales par exemple.

Diapositive 3 (plan suite)

Commençons donc par décrire les données.

Diapositive 4

Les données sur lesquelles on travaille sont constituées par n individus pour lesquels on dispose de deux variables qualitatives. Ces données sont regroupées dans un tableau dit de contingence dans lequel on trouve, en ligne les modalités de la variable V_1 , en colonne les modalités de la variable V_2 , et à l'intersection de la ligne i avec la colonne j on trouve x_{ij} le nombre d'individus ayant choisi ou possédant à la fois la modalité i de V_1 et la modalité j de V_2 . Ce tableau de contingence est aussi appelé tableau croisé dans la terminologie des enquêtes. Donnons quelques exemples de tableaux de contingence.

Un premier exemple, historique parce que c'est sans doute une des premières applications de l'analyse des correspondances, est l'étude du vocabulaire des personnages de la pièce de théâtre Phèdre. On place les personnages en lignes, les mots utilisés par les personnages en colonnes et en x_{ij} le nombre de fois que le personnage i a utilisé le mot j .

Deuxième exemple, on s'intéresse à la perception de parfums de luxe. Pour cela, on a demandé à des dégustateurs d'associer des mots à chacun des parfums. On place les parfums en lignes, les mots en colonnes et en x_{ij} le nombre de fois où le parfum i a été décrit par le mot j .

Enfin en écologie, on s'intéresse à la diversité écologique de plusieurs milieux. On place en lignes les milieux, en colonnes les espèces de plantes ou d'animaux et en x_{ij} le nombre de plantes j trouvées dans le milieu i .

De telles données sont nombreuses et pour ceux qui connaissent le test du χ^2 d'indépendance, il s'agit des mêmes tableaux de données.

Diapositive 5

Les données qui vont nous servir à illustrer notre présentation de l'analyse des correspondances sont les suivantes. Il s'agit d'un tableau qui croise les pays du G8 en lignes et les différentes disciplines dans lesquelles les prix Nobel sont décernés. Dans une case du tableau, on dénombre le nombre de personnes nées dans un des pays du G8 et ayant obtenu un prix dans la discipline concernée. Le tableau donne donc la répartition des 570 prix Nobel distribués entre 1901 et 2015. Par exemple 51 Américains ont obtenu le prix Nobel en chimie. Ce tableau croisé a été complété par ce que l'on appelle des marges. Ici on a ce qu'on appelle la marge colonne, marge colonne parce qu'elle se présente comme une colonne, mais chacun de ces termes, par exemple 80, est une somme qui est faite en ligne. 80 est bien la somme de $24 + 1 + 8 + 18 + 5 + 24$. Il y a aussi une marge ligne ; cette marge ligne s'appelle marge ligne parce qu'elle se présente comme une ligne mais pour l'obtenir on fait la somme des termes en colonne. Ainsi 121 est bien la somme de $4 + 8 + 24 + 1 + 6 + 4 + 23 + 51$.

Donc on se pose la question suivante : Y a-t'il une relation entre les pays et les disciplines ? Autrement dit, est-ce que certains pays sont spécialisés dans certaines disciplines ? Et est-ce que certaines disciplines priment plutôt les ressortissants de certains pays ?

Diapositive 6

Un tableau de contingence est généralement obtenu à partir du croisement de 2 variables. Donc initialement, les données sur lesquelles on travaille sont constituées par n individus pour lesquels on dispose de deux variables qualitatives. Dans le tableau de données que l'on voit ici, l'individu l possède la modalité i de la variable $V1$ et la modalité j de la variable $V2$.

On construit alors un tableau de contingence ou tableau croisé en mettant en regard les modalités de la première variable $V1$, en ligne par exemple, et les modalités de la seconde variable $V2$, ici en colonne. A l'intersection de la ligne i et de la colonne j , on trouve x_{ij} , le nombre de personnes ayant choisi, ou possédant la modalité i de $V1$ et la modalité j de $V2$. Si l'on calcule la somme de tous les termes de ce tableau on trouve n . Ce tableau contient donc la distribution des n individus, dans notre cas, les $n=570$ Nobel qui sont répartis dans les $l \times j$ cases du tableau.

Diapositive 7

A partir de ce tableau de contingence, on va calculer le tableau de probabilités correspondant. Et l'analyse des correspondances va en fait travailler sur le tableau des probabilités. Pour obtenir une probabilité, par exemple f_{ij} , on divise simplement l'effectif x_{ij} par l'effectif total n . On obtient bien ainsi la probabilité conjointe de posséder à la fois la modalité i de $V1$ et la modalité j de $V2$. C'est ce terme que l'on met dans le tableau. Quand on fait la somme de tous ces termes, on obtient 1, ce qui est bien la marque d'une distribution de probabilités.

Ce tableau va être complété par sa marge colonne. On va noter $f_{i.}$ la somme des termes de la ligne i .

De même on va compléter ce tableau par une marge ligne dont le terme général $f.j$ est obtenu en faisant la somme des termes de la colonne j .

Notre problème est de regarder la liaison entre $V1$ et $V2$, c'est-à-dire l'écart entre les données observées et une situation d'indépendance.

Diapositive 8 (plan)

Donnons quelques précisions sur l'indépendance entre 2 variables qualitatives et comment l'analyse des correspondances appréhende l'écart à l'indépendance.

Diapositive 9

Commençons par rappeler le modèle d'indépendance pour 2 événements : 2 événements sont indépendants si la probabilité de A et B est égale au produit des probabilités, probabilité de A fois probabilité de B.

Pour 2 variables qualitatives, on va avoir f_{ij} , probabilité de i et de j , est égale à f_i probabilité de i multipliée par f_j probabilité de j . Mais ceci doit être vrai, bien sûr, quelque soit i et quelque soit j . On dit que la probabilité conjointe est le produit des probabilités marginales.

Il y a une autre écriture de ce modèle d'indépendance que l'on voit ici. f_{ij} sur f_i est égal f_j . Dans ce cas, on dit que la probabilité conditionnelle est égale à la probabilité marginale. Ceci peut paraître un peu formel mais finalement cette écriture est plus proche de l'intuition d'indépendance. La probabilité conditionnelle ici c'est la probabilité d'être j sachant que l'on est i , et donc s'il y a indépendance, cette probabilité est égale à la probabilité d'être j même si l'on ne dispose d'aucune information quant à l'autre variable, c'est-à-dire i . Il y a enfin l'écriture symétrique dans laquelle on divise f_{ij} par f_j et selon le modèle d'indépendance cette nouvelle probabilité conditionnelle, c'est-à-dire la probabilité cette fois d'être i sachant que l'on est j , est bien égale à la probabilité marginale d'être i .

Diapositive 10

La liaison entre 2 variables qualitatives, c'est l'écart entre les données observées d'une part, c'est-à-dire f_{ij} , et le modèle d'indépendance d'autre part, c'est-à-dire f_i que multiplie f_j .

Quand on étudie la liaison entre 2 variables qualitatives, on réalise généralement en premier lieu un test de significativité de la liaison en mettant en œuvre un test du χ^2 . On rappelle ici la forme du test du χ^2 dans lequel on confronte les effectifs observés avec les effectifs théoriques. L'effectif théorique est tout simplement la probabilité selon le modèle d'indépendance, c'est-à-dire $f_i \cdot f_j$ multiplié par l'effectif total n .

Dans cette quantité du χ^2 , on peut mettre n en facteur et faire apparaître la quantité $n \Phi^2$. Le Φ^2 , lui, confronte les probabilités observées et les probabilités théoriques ; c'est donc un indicateur d'intensité de la liaison ; c'est l'écart entre probabilité observée et probabilité théorique. Pourquoi intensité de la liaison ? Parce que ce terme ne dépend pas de l'effectif mais uniquement des probabilités.

Il y a enfin un troisième point qui est la nature de la liaison, c'est-à-dire comment les modalités des deux variables s'associent entre elles.

L'analyse des correspondances travaille sur le tableau des probabilités, elle ne dit donc rien sur la significativité. Son objectif essentiel est de visualiser la nature de la liaison entre les deux variables. Petit

récapitulatif sur ces trois points concernant l'étude d'une liaison. Distinguons bien la significativité ... de l'intensité. Pour l'intensité, par exemple, on a réalisé une petite enquête et on dit qu'à l'issue de cette enquête toutes les personnes qui ont les yeux bleus portent des lunettes. On ne peut pas imaginer de liaison plus forte puisque systématiquement pour la variable couleur des yeux, si l'on sait que l'on a les yeux bleus alors on porte des lunettes. L'intensité de la liaison est très grande. Mais maintenant si je vous dis : en fait, cette enquête a porté sur 4 personnes. Alors, évidemment, on imagine bien, sans même réaliser le test du Khi^2 que ceci n'est pas significatif.

Diapositive 11

Comment l'analyse des correspondances appréhende-t-elle l'écart à l'indépendance ? On va d'abord adopter le point de vue d'une analyse par ligne. Dans cette analyse, on se réfère au modèle d'indépendance suivant : la probabilité conditionnelle sachant i est égale à la probabilité marginale. Pour cela, dans le tableau, on va diviser chaque élément d'une ligne par sa marge. On obtient ainsi ce qu'on appelle un profil ligne i qui n'est rien d'autre qu'une distribution conditionnelle. Dans cette ligne, on a la distribution des réponses pour la variable V_2 sachant que l'on possède la modalité i de la variable V_1 . Ce profil ligne, on va le comparer au profil ligne moyen qui est donc la distribution marginale de la variable V_2 .

Alors pour comparer la distribution conditionnelle à la distribution marginale, l'analyse des correspondances va comparer les profils lignes au profil ligne moyen.

Il s'agit bien d'une approche multidimensionnelle de l'écart à l'indépendance parce que l'on considère simultanément l'ensemble des profils. Donc la question c'est bien de comparer la probabilité conditionnelle f_{ij} sur $f_{i.}$ et la probabilité marginale $f_{.j}$ mais ceci pour l'ensemble des j .

Diapositive 12

Tout ceci est un petit peu formel. Illustrons dans l'exemple ce que signifie : comparer un profil ligne au profil ligne moyen. Voici le tableau des profils de chaque pays complété par le profil ligne moyen. La somme de chaque ligne vaut 100.

Colorions les colonnes, i.e. les disciplines, d'une largeur proportionnelle au pourcentage de la discipline dans le profil moyen, donc par rapport à la ligne du bas.

Construisons pour chaque ligne, donc chaque pays, un graphe en barres empilées afin de visualiser les profils. Examinons la ligne "Italie" ; il s'agit de la distribution du nombre de prix obtenus par catégorie pour les personnes nées en Italie. Plus concrètement, parmi les lauréats du Nobel nés en Italie, 31.6% ont obtenu un prix en littérature. Pour apprécier ce pourcentage, il faut le comparer à 8.6% qui est le pourcentage de prix attribués en littérature pour l'ensemble des pays; donc le pourcentage de prix Nobel en littérature en Italie est particulièrement grand. Ainsi, on va comparer la répartition des prix Nobel en Italie par rapport à la répartition totale des prix Nobel (dans tous les pays).

Cette comparaison du profil ligne "Italie" au profil moyen permet de répondre à la question : les Italiens obtiennent-ils leurs prix plutôt dans certaines disciplines, et sont-ils peu primés dans d'autres disciplines. Autrement dit, sont-ils particulièrement primés dans certaines disciplines ?

Diapositive 13

Nous allons maintenant étudier ce même tableau mais en procédant par colonne. Alors lorsqu'on réalise une analyse par colonne, on se réfère au modèle d'indépendance suivant : f_{ij} sur $f_{.j}$ égale $f_{i.}$. Précisons. On construit le tableau suivant des profils colonnes. Examinons la colonne j . f_{ij} sur $f_{.j}$, c'est la probabilité conditionnelle de répondre i sachant que l'on a déjà répondu j . La somme des termes de cette colonne vaut 1, il s'agit bien d'une distribution de probabilité.

De même que l'on a construit le profil colonne j , on va construire le profil colonne moyen. Ce profil colonne moyen a pour terme général $f_{i.}$ qui est la probabilité de répondre i .

L'idée de l'analyse des correspondances est de comparer les profils colonnes au profil colonne moyen. Il s'agit là encore d'une approche multidimensionnelle de l'écart à l'indépendance, multidimensionnelle parce que l'on va considérer dans l'ensemble tous les termes d'une colonne que l'on va comparer à tous les termes correspondant de la colonne du profil moyen.

Diapositive 14

Examinons dans l'exemple ce que signifie "comparer un profil colonne au profil moyen". Voici les profils colonnes, la somme d'une colonne vaut 100.

On peut colorier chaque ligne en fonction de la valeur prise par la ligne dans le profil colonne moyen.

Construisons pour chaque profil colonne un graphe en barres empilées afin de visualiser les profils. Et comparons chaque profil au profil colonne moyen. Par exemple examinons la colonne "Littérature". La répartition des prix Nobel de littérature par pays semble très différente du profil moyen. Par exemple, le pourcentage de prix Nobel en littérature attribués aux auteurs Américains est plus faible (16.3% en littérature à comparer à 45.1%). En revanche, en littérature, la France et l'Italie sont particulièrement plus primés par rapport à la répartition globale des prix de ces pays.

Donc quand on compare ce profil colonne au profil colonne moyen, on répond à la question suivante : La répartition par pays des prix Nobel en littérature est-elle la même que la répartition de l'ensemble des prix Nobel ?

Nous avons vu sur quelles données travaille l'analyse des correspondances et que l'analyse des correspondances compare une situation à une situation d'indépendance. Nous verrons dans la prochaine vidéo comment visualiser les écarts à l'indépendance en construisant un nuage des lignes et un nuage des colonnes.

Deuxième partie. Visualisation des nuages des lignes et des colonnes

(Diapositives 15 à 21)

Diapositive 15 (plan)

Nous avons vu que l'analyse des correspondances allait travailler à partir du modèle d'indépendance. Pour présenter l'analyse des correspondances, nous allons maintenant raisonner essentiellement géométriquement et construire deux nuages de points.

Diapositive 16

Tout d'abord nous allons construire le nuage de profils lignes. Un profil ligne est un ensemble de J valeurs numériques, c'est donc un point dans l'espace à J dimensions que l'on note R^J ; chaque dimension de cet espace correspondant à une modalité j de la variable V_2 . Ainsi, sur la dimension j , le profil ligne i a comme coordonnée f_{ij} sur $f_{i.}$, c'est-à-dire le j ème terme de son profil. Si l'on considère l'ensemble des profils lignes i , on obtient le nuage des profils lignes ce que l'on note N_I .

Dans ce nuage on peut situer ce que l'on a appelé le profil moyen qui a comme coordonnée pour la j ème dimension $f_{.j}$. Ce profil moyen on l'appelle G comme centre de gravité. G peut en effet être considéré comme le centre de gravité du nuage N_I à condition d'affecter, à chaque point i , un poids proportionnel à son effectif marginal, donc à $f_{i.}$. Dans cet espace, ce qui nous intéresse surtout, c'est de comparer la position du profil ligne i au profil moyen. Pour cela, on va placer l'origine du nuage au profil moyen.

Dans cet espace, il faut savoir calculer une distance. La distance qui est utilisée dans l'analyse des correspondances est appelée distance du Chi^2 . Elle ressemble beaucoup à la distance euclidienne usuelle. En effet, on va retrouver une somme des carrés des écarts. Ici on a l'écart entre le profil i et le profil i' , f_{ij} sur $f_{i.}$ et $f_{i'j}$ sur $f_{i'.$; on élève au carré et on fait la somme sur tous les j . La différence avec la distance euclidienne usuelle, c'est que chaque dimension j a un poids qui est $1/f_{.j}$. La justification de cette distance apparaîtra dans la suite du cours.

Le centre de gravité du nuage, qui correspond au profil moyen, a pour coordonnée $f_{.j}$, donc la distance entre le profil i et le profil moyen est simplement cette distance.

Diapositive 17

De même que l'on a construit un nuage des profils lignes, on va construire un nuage des profils colonnes. Un profil colonne, c'est un ensemble de I valeurs numériques, c'est donc un point dans un espace à I dimensions. Dans cet espace, chaque dimension correspond à une modalité i de V_1 . Le long de cette modalité i , la coordonnée du profil colonne est $f_{ij}/f_{.j}$. Il y a J profils colonnes qui constituent à eux tous le nuage N_J .

A ce nuage, on ajoute le profil moyen G_J qui a donc comme coordonnées $f_{.i}$ sur l'axe i . Ce profil moyen peut être considéré comme le centre de gravité du nuage à la condition d'affecter à chaque profil colonne j un poids égal à son effectif marginal, ou plus exactement à sa probabilité marginale, c'est-à-dire ici $f_{.j}$. Dans cet espace, on accorde une importance très grande à la distance entre un profil colonne et le profil moyen. Pour cela, on va mettre l'origine au centre de gravité.

Pour calculer la distance dans cet espace, on utilise, comme dans le nuage des profils lignes, la distance du χ^2 . Ici cette distance du χ^2 s'écrit de la façon suivante : c'est une somme des carrés des différences des coordonnées mais chaque différence des coordonnées est pondérée par $1/f_i$.

Le centre de gravité du nuage, qui correspond au profil moyen, a pour coordonnée f_i , donc la distance entre le profil j et le profil moyen est simplement la racine carrée de cette quantité.

Diapositive 18

Alors que se passe-t-il s'il y a indépendance ? S'il y a indépendance, la probabilité conditionnelle est égale à la probabilité marginale, c'est-à-dire que dans chacun des nuages, tous les profils sont confondus avec le profil moyen. Autrement dit, le nuage dans ce cas est réduit à un seul point, l'origine des axes.

Diapositive 19

Donc, plus les données s'écartent de l'indépendance et plus les profils s'écartent de l'origine.

Calculons l'inertie du nuage N_I par rapport à son centre de gravité. L'inertie d'un nuage de points c'est la somme des inerties de ces points. L'inertie d'un point c'est la masse par le carré de distance donc on a bien ici l'inertie du point i qui vaut f_i multiplié par le carré de la distance du χ^2 entre i et G_I . Lorsque l'on effectue les calculs, on obtient le résultat intermédiaire suivant qui n'est autre que le Φ^2 ou si l'on préfère le χ^2 sur n . Ainsi l'inertie du nuage de points est bien un indicateur d'intensité de l'écart à l'indépendance.

Comme ce qui nous intéresse c'est d'étudier l'écart à l'indépendance, ceci revient à étudier l'inertie de N_I . Tel est le point de vue de l'analyse des correspondances. On peut faire le même raisonnement pour le nuage N_J et on a le résultat suivant : l'inertie pour le nuage N_I est égale à l'inertie du nuage N_J . Ça c'est un résultat capital qui est englobé dans ce qu'on appelle la dualité, c'est-à-dire le caractère double. En fait il revient au même d'analyser des tableaux en termes de lignes ou en termes de colonnes. Ceci est très important en analyse des correspondances. On s'aperçoit que, en analyse des correspondances, lignes et colonnes jouent des rôles parfaitement symétriques. C'est là une propriété très importante qui la distingue d'autres méthodes comme l'analyse en composantes principales par exemple.

Diapositive 20

Nous avons traduit notre question initiale qui est celle de la liaison entre deux variables qualitatives comme d'abord un écart à l'indépendance. Cet écart à la situation d'indépendance s'interprète géométriquement comme une inertie du nuage N_I par rapport à l'origine des axes. Pour étudier cette inertie, comment procède l'analyse des correspondances ? Et bien elle décompose cette inertie par analyse factorielle. Ce qu'on appelle ici analyse factorielle c'est le dénominateur commun à toutes les méthodes dites factorielle, à savoir l'analyse en composantes principales, l'analyse des correspondances bien sûr, l'analyse des correspondances multiples enfin. Et donc l'analyse factorielle procède de la façon suivante : elle projette un nuage de points sur une suite d'axes orthogonaux d'inertie maximum.

Quand on a une suite d'axes, on peut apparier les axes et constituer ainsi des plans. Examinons comment on va trouver le premier plan. Alors la meilleure représentation plane du nuage N_I s'obtient, comme on l'a dit, par projection donc sur un plan P . On a ici le point M_i qui correspond au profil de i . Ce point M_i est projeté sur le plan en H_i . En fait, ce que l'utilisateur regarde ce sont les points H_i et il espère que le point H_i n'est pas trop différent du point M_i pour pouvoir conclure. Quel est le critère que l'on utilise pour trouver le plan P ?

Ce que l'on cherche à bien représenter, rappelons-le, c'est l'inertie du nuage N_I et ce que l'on cherche à maximiser c'est l'inertie projetée. Examinons cette relation. On reconnaît ici l'inertie projetée du point i , c'est-à-dire la masse du point i , f_i , par le carré de sa distance à l'origine. En intégrant le signe somme, on a bien l'inertie projetée du nuage N_I . Donc la question est de trouver le plan P qui rend maximum l'inertie projetée. On dit que l'on a un plan d'inertie maximum. Alors pour trouver ce plan, on va d'abord chercher un axe u_1 , d'inertie maximum, puis un second axe u_2 d'inertie maximum avec la contrainte que u_2 doit être orthogonal à u_1 . On a donc apparié u_1 et u_2 pour avoir le plan P . L'inertie d'un axe s sera notée λ_s , λ_s parce qu'il s'agit d'une valeur propre, s parce qu'il s'agit de la valeur propre de rang s , les valeurs propres étant rangées par ordre décroissant.

Diapositive 21

Voici le graphique fourni par l'analyse des correspondances quand on l'applique à notre tableau de données. On a donc un graphique plan sur lequel se trouvent et les lignes, et les colonnes. Pour les lignes, ce sont les points en bleu, à savoir les pays. Pour les colonnes, on a les points rouges à savoir les catégories de prix. Quelles sont les règles d'interprétation ? Un point sur lequel nous avons beaucoup insisté c'est la distance au centre de gravité. Examinons par exemple le point du Royaume-Uni. Il est très proche du centre de gravité.

Regardons le profil du Royaume-Uni : 25%, 6%, 8%, 28%, 12% et 22%. Et ce profil, comparons-le maintenant au profil moyen : on voit qu'il est très proche 25 % contre 21%, 6% contre 11%, 8% contre 9%, 28% contre 25%, 12% contre 9% et 22% contre 26%. On voit bien que le profil du Royaume-Uni est proche du profil moyen ce qui se concrétise par une proximité entre le point correspondant et l'origine. Si maintenant on choisit un autre point, par exemple l'Italie. Et bien l'Italie est assez loin de l'origine ce qui veut dire que le profil de l'Italie est différent du profil moyen. Examinons le profil de ce pays : par exemple 5% des prix obtenus par les Italiens le sont en chimie alors que dans la population le pourcentage est de 21%. En revanche, 32% des prix italiens sont obtenus en littérature alors que 9% des prix sont obtenus en littérature pour l'ensemble des pays du G8. Un point important pour l'utilisateur est de nommer un axe. Alors comment nommer un axe ici ? Considérons les éléments d'un ensemble, par exemple les points bleus, c'est-à-dire les lignes donc les pays. On trouve des oppositions entre les pays d'Amérique du nord sur la gauche, les pays latins (France et Italie) sur la droite et Japon et Allemagne dans le bas du graphe. Il est difficile de décrire les axes avec les pays sauf si on sait par exemple que les Nord-Américains sont réputés en sciences économiques et que Français et Italiens sont plus tournés vers la littérature. Maintenant si l'on regarde les points rouges, c'est-à-dire les catégories de prix, ces points rouges se répartissent avec sur la gauche les prix obtenus en médecine et sciences économiques et sur la droite les prix Nobel de littérature et de la paix. Le deuxième axe, l'axe vertical, oppose quant à lui les prix Nobel de physique et chimie au prix Nobel en sciences économiques.

On peut penser que le 1^{er} axe oppose les prix scientifiques aux autres prix, tandis que le deuxième axe oppose la physique - chimie aux sciences économiques. L'interprétation des axes est la même entre les lignes et les colonnes en raison de ce que nous avons déjà appelé la dualité. En effet, on étudie le même tableau au travers des lignes et au travers des colonnes mais c'est bien le même tableau que l'on analyse. Le tableau des données confronté à ce que l'on obtient si l'on avait exactement l'indépendance.

Nous avons vu comment visualiser le nuage des lignes et le nuage des colonnes. Nous verrons dans les prochaines vidéos deux particularités de l'analyse des correspondances : l'inertie et la représentation simultanée.

Troisième partie. Inertie et pourcentage d'inertie

(Diapositives 22 à 26)

Diapositive 22 (plan)

Nous avons vu comment construire le nuage des lignes et le nuage des colonnes et comment projeter le nuage des lignes et le nuage des colonnes pour obtenir des représentations graphiques. En analyse des correspondances, comme dans toute analyse factorielle, les premiers indicateurs que l'on regarde sont les pourcentages d'inertie. Nous allons voir aussi dans cette section qu'en analyse des correspondances, les inerties sont également très particulières.

Diapositive 23

Commençons par commenter les pourcentages d'inertie. Comme dans toute analyse factorielle, les pourcentages d'inertie sont les premiers indicateurs que l'on regarde. La question posée est la suivante. On a une représentation du nuage, quelle est la qualité de cette représentation ? De façon générale, la qualité de représentation est mesurée par le rapport inertie projetée sur inertie totale. Généralement ce rapport est multiplié par 100 pour l'exprimer en pourcentage. Dans le cas particulier du nuage NI et de l'axe de rang s , la qualité de représentation du nuage NI sur l'axe de rang s sera donc l'inertie projetée de NI sur U_s divisé par l'inertie totale de NI. Ceci peut s'écrire avec les notations utilisées précédemment $\lambda_{s,k}$ sur la somme des $\lambda_{k,k}$. Généralement ceci, comme il a été dit, est exprimé en pourcentage.

Examinons les pourcentages d'inertie de l'exemple. On voit que pour le premier axe, le pourcentage d'inertie est de 54.75, ce qui est important. On pourra donc dire que le 1er axe représente 54.75% de l'écart à l'indépendance. Pour le deuxième axe, le pourcentage d'inertie est de 24.60% et donc les deux premiers axes résument à eux deux près de 79% de l'écart à l'indépendance et donc on pourra se limiter à ces deux axes dans l'interprétation.

Propriété : les inerties projetées, c'est-à-dire les valeurs propres s'additionnent d'un axe à l'autre. Ceci tient au fait que les axes sont orthogonaux. Donc la somme de toutes les valeurs propres, de toutes les inerties projetées si on préfère, est égale à l'inertie totale du nuage NI, ce qui est vrai pour toutes les analyses factorielles. Dans le cas particulier de l'analyse des correspondances cette inertie totale est égale au Φ^2 . On peut faire un petit calcul dans le cas de l'analyse des correspondances sur notre jeu de données exemple : on multiplie l'inertie totale de 0.1522 par n , somme du tableau qui vaut 570 et on tombe sur un χ^2 qui vaut 86.75. Compte tenu du nombre de degrés de liberté, la probabilité critique est de 2.77 fois 10 puissance moins 6 donc extrêmement petite, la significativité de la liaison entre les pays et les catégories est bien entendue hors de doute dans cet exemple.

A ces 2 premiers points concernant les pourcentages d'inertie, s'en ajoute un 3ème: la décroissance des inerties en fonction du rang s des axes suggère le nombre d'axes à conserver. Nous montrons ici la décroissance des valeurs propres d'une AFC effectuée sur un tableau de contingence croisant 10 vins blancs du Val de Loire décrit par 30 mots. Les 10 vins blancs du Val de Loire sont en ligne, les mots sont en colonne et x_{ij} est le nombre de fois que le mot j a été associé au vin i . Quand on regarde la séquence des valeurs propres avec un diagramme en barres, on observe que les 2 premières valeurs propres sont sensiblement plus grandes que les suivantes. Les 2 premiers axes sont donc prépondérants du point de vue de l'inertie ce qui suggère d'interpréter de façon privilégiée le plan constitué par ces 2 premiers axes.

Diapositive 24

En analyse des correspondances, il convient de bien distinguer les inerties des pourcentages d'inertie car les inerties sont très particulières. Elles constituent une composante du Φ^2 qui est cet indicateur de liaison global entre 2 variables mises en correspondance. En analyse des correspondances le résultat théorique suivant est très important : les valeurs propres sont toujours comprises entre 0 et 1. Rappelons qu'en analyse en composantes principales quand les variables sont normées, il en va différemment puisque la première valeur propre, elle, est automatiquement supérieure ou égale à 1.

Que signifie en analyse des correspondances une valeur propre égale à 1. Ce cas limite est tout à fait intéressant. Et à quelle structure de données correspond-il ? Voici la structure à laquelle il correspond.

On peut séparer les lignes en 2 blocs I1 et I2. Les colonnes peuvent être aussi séparées en 2 blocs J1 et J2. Et cette double partition est le lieu d'une association exclusive c'est-à-dire que les lignes du bloc I1 s'associent uniquement à J1 et absolument pas à J2 et les lignes du bloc I2 s'associent exclusivement à J2 et absolument pas à J1. C'est bien la marque d'une liaison très forte puisque l'on a une association exclusive entre les modalités d'une variable et les modalités de l'autre. Du point de vue graphique, on va obtenir le graphique suivant. L'axe correspondant à une valeur propre égale à 1 oppose parfaitement le bloc I1 et I2 c'est-à-dire qu'à l'intérieur de I1 on ne fait aucune distinction et il oppose parfaitement J1 et J2. A l'intérieur de J1, on ne fait aucune distinction.

Diapositive 25

Ces inerties qui sont très particulières, on va les regarder sur un autre jeu de données. Il s'agit de données concernant la reconnaissance de trois saveurs, le sucré, l'acide, l'amer. Le plan l'expérience est le suivant : pour chaque saveur on a demandé à 10 personnes de reconnaître la saveur d'une solution qui leur était présentée. Voici le petit tableau de données. Lisons-le par ligne : la solution sucrée a été perçue 10 fois comme sucrée et jamais comme acide, jamais comme amer. La solution acide n'a jamais été perçue comme sucrée par contre elle a été perçue 9 fois comme acide et 1 fois comme amer. Lorsqu'on réalise l'analyse des correspondances de ce tableau, on obtient une première valeur propre de 1 qui est la marque de la structure bloc-diagonal que nous avons déjà citée. Effectivement, reportons-nous au tableau : quand on regarde la première ligne et la première colonne, on a bien le sucré qui est perçu uniquement comme sucré et la perception sucrée (là je m'intéresse à la colonne) n'est associée qu'au sucré. On a donc une valeur propre de 1. Le graphique correspondant oppose donc parfaitement d'un côté sucré et perçu sucré, ces deux points sont confondus, à de l'autre côté amer, perçu amer, acide et perçu acide qui sont parfaitement confondus. Examinons maintenant le 2ème axe de cette analyse des correspondances. Ce 2ème axe oppose d'un côté amer et perçu amer à de l'autre, acide et perçu acide. Il exprime que, dans l'ensemble, amer est plutôt perçu comme amer et acide est plutôt perçu comme acide. C'est bien ce qui se passe. Néanmoins, quand on regarde sur le tableau il y a une confusion : acide n'est pas toujours perçu comme acide, amer n'est pas toujours perçu comme amer. Comment pouvons-nous déceler cette confusion ? Tout simplement par la valeur propre. Si on n'avait aucune confusion, on aurait une valeur propre de 1 or la valeur propre vaut seulement 0.375. Cette valeur propre de 0.375, beaucoup plus petite que 1, nous indique qu'il y a une confusion entre amer et acide, c'est-à-dire que, de temps en temps, amer est perçu acide et acide est perçu comme amer. Sur ce graphique cette confusion est parfaitement perceptible parce que, sur le premier axe, on a une situation de référence avec une situation parfaite sans confusion et on voit bien que l'opposition sur le premier axe entre sucré d'une part et amer-acide d'autre part, correspondent à une inertie beaucoup

plus grande que l'opposition entre acide et amer. Acide et amer sont beaucoup plus proches entre eux qu'ils ne sont proches de sucré. Donc sur ce graphique on voit bien qu'il y a une plus grande confusion entre acide et amer qu'entre acide-amer d'une part et sucré d'autre part.

On peut examiner un autre tableau dans lequel on a augmenté la confusion. C'est-à-dire que cette fois, dans ce tableau, l'acidité a été perçue seulement 7 fois comme acide et non pas 9 fois comme précédemment et l'amertume a été perçue que cinq fois comme amer. Examinons les résultats de cette analyse : on trouve toujours une première valeur propre de 1, qui nous indique que le sucré est parfaitement reconnu par contre la valeur propre de l'axe 2, cette fois, n'est plus de 0.375 mais de 0.04. Examinons le graphique : le graphique ressemble tout à fait au précédent c'est-à-dire que le premier axe montre le rôle particulier du sucré et le deuxième axe oppose acide et amer. Simplement l'écart entre acide et amer est beaucoup plus petit que dans le cas précédent. Et c'est cette opposition entre acide et amer qui se marque par une distance plus petite qui est bien la conséquence de la plus grande confusion entre acide et amer. Finalement cette valeur propre de 0.04 associée au 2ème axe nous indique le degré de confusion entre acide et amer. Et dans le cas extrême où on n'aurait aucune confusion (avec un tableau diagonal), on aurait 2 valeurs propres égales à 1. Quel bilan pouvons-nous tirer de ces deux exemples ? On s'aperçoit que, dans chacun des exemples, le deuxième axe est toujours le même : il oppose d'une part amer et perçu amer à d'autre part acide et perçu acide. On peut dire que dans ces deux cas le 2ème axe est pratiquement le même. Il a la même signification, il indique simplement qu'amer est globalement plutôt perçu amer et acide est globalement plutôt perçu acide. Maintenant, d'un exemple à l'autre, ce globalement n'a pas du tout la même force puisque dans le premier cas on peut dire que la reconnaissance est bonne ; dans le deuxième cas elle est plutôt médiocre. Cela nous montre que le graphique nous indique l'opposition entre acide et amer mais il ne nous dit rien quant à la force de cette opposition. C'est la valeur propre qui va nous dire si cette opposition est très forte ou pas. Si elle vaut 1, cette opposition est tout à fait drastique (si l'on peut dire), c'est-à-dire qu'elle est le lieu d'une association exclusive. Si par contre cette valeur propre est assez faible, c'est le cas du deuxième exemple dans lequel cette valeur propre valait 0.04 ; et bien dans ce cas-là on a à peine une petite reconnaissance majoritaire puisque l'amer n'est reconnu que dans la moitié des cas et l'acide dans 7 cas sur 10. Cela signifie qu'il faut toujours commencer par regarder les valeurs propres en analyse des correspondances car ces valeurs propres nous indiquent si les associations que l'on va mettre en évidence sont très nettes ou sont peu nettes dans les données. Le graphique ne dit rien à ce sujet puisqu'il ne parle pas d'intensité de liaison mais nous indique simplement la nature de la liaison. Ici la nature de la liaison c'est acide s'associe plutôt à perçu acide et amer s'associe plutôt à perçu amer.

Diapositive 26

Revenons aux données sur les prix Nobel. Nous avons déjà commenté les pourcentages d'inertie de 55% et 25% en disant qu'ils étaient importants et qu'ils nous incitaient à conserver uniquement les 2 premiers axes dans l'interprétation. Les inerties elles-mêmes sont de 0.083 et 0.037, ce qui peut être considéré comme faible, puisque très sensiblement inférieure à 1. Cette valeur 1 correspondrait à une association exclusive entre modalité, par exemple entre un pays et une catégorie. On est donc très loin de cette association exclusive. Effectivement, les prix Nobel sont distribués à des personnes de toutes les nationalités. Si l'on regarde maintenant la somme des inerties, comme on sait qu'en analyse des correspondances, cette somme correspond aux Φ^2 , cette somme vaudrait au maximum 5 si chacune des inerties valait 1, on est donc ici encore, bien entendu, très loin du maximum. Dans ce tableau, on peut dire que l'on est très très loin d'une association exclusive entre les modalités des 2 variables.

Quatrième partie. La représentation simultanée

(Diapositives 27 à 33)

Diapositive 27 (plan)

Alors revenons maintenant sur l'interprétation du graphe de l'AFC afin d'introduire une caractéristique essentielle en analyse des correspondances. Pour l'instant nous avons commenté le graphe en considérant d'une part les lignes, et d'autre part les colonnes. L'analyse des correspondances permet une représentation simultanée des lignes et des colonnes. Dans son principe cette représentation simultanée est possible car lignes et colonnes sont des éléments de même nature : il s'agit de modalités de variables qualitatives.

Diapositive 28

La représentation simultanée des lignes et des colonnes fonctionne en analyse des correspondances grâce à des relations qu'on appelle relations de transition ou encore propriétés barycentriques. Voici ici cette relation de transition concernant l'axe de rang s . Précisons les différents termes de cette relation. $F_s(i)$ est la coordonnée de la ligne i le long de l'axe de rang s . $G_s(j)$ est la coordonnée de la colonne j le long de l'axe de rang s . Donc on voit que l'on sait relier les coordonnées des lignes aux coordonnées des colonnes. Plus précisément, comme il y a un signe somme sur les j , on sait exprimer la coordonnée d'une ligne par rapport aux coordonnées de toutes les colonnes. f_{ij} sur $f_{i.}$, c'est le j ème terme du profil. Donc on est en train de faire une somme sur les j , donc une somme sur toutes les colonnes, des coordonnées des colonnes pondérées par les éléments du profil i .

Donc on fait en quelque sorte une moyenne des coordonnées des colonnes et dans cette moyenne, une colonne j intervient d'autant plus qu'elle a une forte probabilité conditionnelle pour le profil i .

Une fois qu'on a calculé ces centres de gravité, ces barycentres si on veut, on les dilate puisque le coefficient $1/\text{racine}(\lambda_s)$ est, par construction, plus grand que 1. On peut formuler cette relation de transition de la façon suivante : la ligne i est au barycentre de toutes les colonnes, chaque colonne étant affecté du poids f_{ij} sur $f_{i.}$ qui est le j ème terme du profil ligne i . Cette propriété barycentrique, concrètement, on va la traduire ainsi de façon simplifiée : une ligne est du côté des colonnes auxquelles elle s'associe de plus. C'est-à-dire que si f_{ij} sur $f_{i.}$ est très grand, alors $G_s(j)$ compte beaucoup dans le calcul de la coordonnée de i .

En analyse des correspondances, lignes et colonnes jouent des rôles symétriques donc si on permute les rôles joués par les lignes et colonnes, on obtient la seconde propriété barycentrique : concrètement, une colonne est du côté des lignes auxquelles elle s'associe le plus. Cette double propriété barycentrique (une ligne est du côté des colonnes auxquelles elle s'associe le plus et une colonne est du côté des lignes auxquelles elle s'associe le plus), cette double propriété barycentrique permet une utilisation de la représentation simultanée, et c'est ce qui en fait sa richesse.

Diapositive 29

Appliquons maintenant ces propriétés barycentriques à un jeu de données que nous avons déjà utilisé : celui de la reconnaissance de saveurs fondamentales. Nous considérons ici 2 réalisations du jeu de données : l'un dans lequel il y a peu de confusion entre acide et amer puisqu'on a $3+1 = 4$ confusions sur 20 données, et un autre cas dans lequel il y a plus de confusions $5 + 3 = 8$ confusions, c'est-à-dire 40% de confusions. 20% de

confusions à gauche, 40% à droite. Les résultats de l'analyse des correspondances de ces 2 tableaux, on les a déjà commentés : la 1ère valeur propre vaut 1, c'est la conséquence de l'association exclusive entre sucré et perçu sucré. Alors quelles conséquences sur les propriétés barycentriques ?

Rappelons la propriété barycentrique : si λ_s vaut 1, cela veut dire qu'une colonne va être à l'exact barycentre des lignes. Alors prenons par exemple perçu sucré : perçu sucré est rigoureusement confondu avec sucré parce qu'il ne s'associe qu'avec sucré. De l'autre côté, on a, le long du premier axe, amer, acide, perçu amer, perçu acide, qui ont exactement la même coordonnée ; perçu amer est donc bien à l'exact barycentre de sucré, amer et acide puisqu'il ne s'associe jamais avec sucré. Examinons maintenant le 2ème axe : le 2ème axe met en évidence une association privilégiée entre amer et perçu amer d'une part et acide et perçu acide d'autre part, et cela dans les 2 cas. La disposition relative de ces 4 points est rigoureusement identique entre la situation de gauche et la situation de droite, ce qui nous montre que cette disposition relative des points nous donne des informations sur la nature de la liaison : quelles sont les modalités qui s'associent entre elles, mais ne nous dit rien sur l'intensité de la liaison. Est-ce que cette association est forte, très forte, voire quasiment exclusive. Si l'on veut avoir des informations sur l'intensité de la liaison, alors il faut regarder les valeurs propres qui sont relativement élevées dans un cas, 0.375, et très faible dans l'autre, 0.042. La plus grande valeur propre indique qu'on a une liaison beaucoup plus forte. On est cependant loin d'une association exclusive, auquel cas on aurait une valeur propre de 1.

Diapositive 30

Analysons maintenant de façon précise comment fonctionne la propriété barycentrique en effectuant un grossissement des coordonnées le long du 2ème axe concernant les points acide et amer. Dans cette propriété barycentrique, on calcule d'abord le barycentre et une fois qu'on a les barycentres, on les dilate avec le coefficient $1/\sqrt{\lambda_s}$.

Prenons comme exemple la construction du point perçu amer. Perçu amer, dans un 1er temps va être le barycentre des points acide et amer avec les coefficients 1 et 7. Donc si l'on calcule le barycentre d'acide et amer avec les coefficients $1/8$ et $7/8$, on tombe sur le point rouge qui est beaucoup plus proche du point amer que du point acide. Effectuons maintenant le même calcul dans le cas de droite ; on va calculer le barycentre entre acide et amer avec cette fois les coefficients 3 et 5. On obtient un rond rouge qui est cette fois du côté d'amer mais assez légèrement dans le rapport 3 et 5. On voit que, si l'on regarde les barycentres, on a bien l'idée de l'intensité de la liaison.

Que fait l'analyse des correspondances ensuite ? Elle multiplie ces coordonnées par $1/\sqrt{\lambda_s}$. Cette valeur, $1/\sqrt{\lambda_s}$, est très différente d'un cas à l'autre puisqu'elle vaut 1.6 dans le 1er cas et 4.9 dans le second cas. Autrement dit, le nuage des barycentres va être beaucoup plus dilaté (dans le cas de droite) lorsqu'on a une faible liaison que lorsque l'on a une forte liaison.

De cette façon, l'analyse des correspondances gomme l'effet de l'intensité de liaison pour ne garder que la nature de la liaison. On peut se demander pourquoi on fait ceci. La première raison évidente, c'est qu'on a exactement la même relation entre une ligne et l'ensemble des colonnes et une colonne par rapport à l'ensemble des lignes. On a bien ici le rôle symétrique joué par les lignes et les colonnes. Une autre façon de voir les choses est de dire : si on regarde les barycentres, dans le cas de liaison faible, on va avoir beaucoup de mal à déceler des associations. Aussi l'analyse des correspondances va grossir le graphique, un peu à la façon d'un microscope, pour mettre en évidence les associations. En conclusion, on peut dire que l'analyse des correspondances ne dit rien sur la significativité de la liaison, elle travaille sur les probabilités (on l'a dit),

l'intensité de la liaison, elle est visible dans les valeurs propres et la nature de liaison dans les positions relatives des points.

Diapositive 31

Utilisons la double propriété barycentrique sur le jeu de données des prix Nobel. Rappelons tout d'abord que le centre de gravité du nuage est au barycentre des points rouges avec la pondération du profil moyen. On peut donc noter sous chaque catégorie le poids de la catégorie qui conduit le profil moyen au barycentre du nuage. Considérons maintenant le Japon. Le Japon correspond à un profil ayant pour terme 26, 0, 9, 13, 4 et 48. C'est donc le barycentre des points rouges affecté de ces nombres.

Si on utilise une police proportionnelle au poids de la catégorie, on remarque que le Japon a un poids fort pour la physique et la chimie par rapport aux poids de ces 2 catégories dans le profil moyen. Le Japon va donc être positionné au barycentre des catégories affectées de ces poids, ce qui va conduire le Japon vers le bas du graphe, du côté de la physique et de la chimie. C'est bien ceci la propriété barycentrique : un point ligne va du côté des colonnes auquel il s'associe le plus (et à l'opposé des colonnes auxquels il s'associe le moins).

Si maintenant on regarde l'Italie, on a le résultat suivant : l'Italie s'associe beaucoup plus à la littérature qu'aux sciences économiques d'où le point de l'Italie du côté de la littérature. On peut noter que les pourcentages de l'Italie pour la physique et la médecine sont élevés, mais ces pourcentages sont en réalité du même ordre de grandeur que ceux du profil moyen. Ainsi, ces deux pourcentages ne contribuent pas à éloigner l'Italie du barycentre du nuage.

Diapositive 32

Grâce aux propriétés barycentriques, nous pouvons donc maintenant interpréter simultanément les points rouges et les points bleus, c'est-à-dire les lignes et les colonnes. Nous avons dit que le premier axe oppose sur la gauche les disciplines scientifiques aux autres disciplines, et on peut donc dire que les USA obtiennent principalement des prix dans les disciplines scientifiques tandis que Français et Italiens obtiennent surtout des prix Nobel de la paix et de littérature. La position plus extrême de l'Italie par rapport à la France montre que les Italiens sont encore plus spécialisés en littérature que les Français. Quantitativement, ils obtiennent moins de prix littéraires que les français mais le prix littéraire est une spécificité en Italie car relativement aux autres catégories, les Italiens obtiennent principalement leur prix Nobel en littérature.

Le second axe oppose en bas du graphe la physique et la chimie, et les Allemands et Japonais obtiennent principalement des prix dans ces disciplines.

Cinquième partie. Les aides à l'interprétation

(Diapositives 33 à 42)

Diapositive 33 (plan)

Examinons maintenant les aides à l'interprétation classiques que sont la qualité de représentation et la contribution. Ces aides sont communes aux différentes méthodes d'analyses factorielles.

Diapositive 34

La qualité de représentation d'un point se mesure à partir du même indicateur que pour un nuage, c'est-à-dire l'inertie projetée du point divisée par l'inertie totale du point. Cet indicateur montre dans quelle mesure l'écart du profil au profil moyen est complètement représenté par l'axe. Reprenons l'interprétation géométrique : le point M_i , qui représente le profil i , est projeté sur U_s en un point His . Pour calculer la qualité de représentation du profil i par l'axe de rang s , on calcule donc le rapport : inertie projetée de M_i sur U_s divisée par l'inertie totale de M_i , ce qui donne $fi \cdot OHis^2 / OMi^2$. Les fi s'éliminent, cela signifie que la qualité de représentation ne dépend pas du poids. Ce que l'on voit, c'est que ce rapport n'est rien d'autre que le cosinus carré de l'angle entre OM_i d'une part et U_s d'autre part. Cette interprétation en termes de cosinus correspond bien à la question : dans quelle mesure l'écart entre le profil et le profil moyen est bien représenté par l'axe. Est-ce que le point M_i va bien dans la même direction que l'axe.

Diapositive 35

Examinons les qualités de représentation sur le petit jeu de données des saveurs dans lequel il y a une confusion importante entre acide et amer. Examinons le point acide : acide a une qualité de représentation de 0.89 sur le 1er axe et 0.11 sur le second. Cela signifie que le point acide s'écarte du point moyen, beaucoup plus le long de l'axe 1 que le long de l'axe 2. Effectivement, le long de l'axe 1, on voit qu'acide ne s'associe jamais avec perçu sucré alors que le long de l'axe 2, acide s'associe quelques fois à perçu amer. Donc l'écart au profil moyen est beaucoup plus exprimé sur le premier axe que sur le second.

En pratique, la qualité de représentation est utilisée de la manière suivante : lorsqu'on a beaucoup de points, pour commencer une interprétation, on sélectionne quelques points qui ont à la fois des coordonnées remarquables, c'est-à-dire qui sont éloignés le long de l'axe que l'on étudie, et à la fois qui ont une bonne qualité de représentation. Ainsi on sélectionne des points qui vont nous aider à bien interpréter puisqu'il sera facile de retrouver dans les données la signification de l'axe sachant que l'essentiel de l'écart entre le profil étudié et le profil moyen s'exprime sur l'axe.

Diapositive 36

La seconde aide à l'interprétation classique qui est commune à toutes les méthodes factorielles est la notion de contribution. Pour calculer la contribution d'un point i à l'inertie d'un axe s , on calcule d'abord un indicateur brut qui est l'inertie projetée du point, soit avec nos notations fi que multiplie $OHis$ au carré.

Cet indicateur brut est ramené en pourcentage en divisant par l'inertie totale de l'axe que nous avons déjà notée λ_s . La multiplication par 100 de ce rapport permet une expression commode en pourcentage.

On peut additionner les contributions de plusieurs éléments, c'est-à-dire calculer la contribution non pas d'un élément mais de 2 ou 3 éléments à la construction d'un axe.

A quoi servent les contributions ? Elles indiquent dans quelle mesure on peut considérer qu'un axe est dû à un élément ou à quelques éléments. Si par exemple, pour un axe donné, on retrouve qu'un élément contribue à 90% à l'inertie de cet axe, on pourra limiter l'interprétation de cet axe à cet élément. Quand une contribution est-elle grande ?

Elle est grande si à la fois la distance OH_i est assez grande et si la masse f_i est assez grande. En ce sens on peut dire que la contribution réalise un compromis opérationnel entre distance à l'origine et poids.

Compromis opérationnel, c'est-à-dire que les contributions sont généralement utilisées dans le cas de grands tableaux pour sélectionner un sous ensemble d'éléments pour commencer l'interprétation. On va ainsi sélectionner un ensemble d'éléments très contributifs. L'idéal étant de trouver un ensemble d'éléments qui soit à la fois très contributifs et bien représentés.

Diapositive 37

Illustrons la notion de contribution sur un exemple. Pour cela nous n'utilisons pas les exemples précédents car dans ces exemples précédents les effectifs marginaux des lignes comme des colonnes sont très peu différents voire pas du tout différents. Nous avons construit ici un petit tableau dans lequel les effectifs marginaux des lignes sont très différents. On remarque ici que les effectifs marginaux de A et de D sont plus de 10 fois plus petits que les effectifs marginaux de B et de C. On réalise l'analyse des correspondances de ce tableau et on obtient les résultats suivants. Le premier axe est prépondérant avec 83 % de l'inertie et nous allons nous y limiter. Il oppose les colonnes X1 à X4, X1 étant caractérisée par les lignes A et B, X4 étant caractérisée par les lignes C et D. Le long de cet axe, les coordonnées de A et de D sont extrêmes par rapport à B et C, elles sont beaucoup plus grandes. Examinons les contributions : quand on regarde ce tableau de contributions, on voit que les contributions de B et C sont plus importantes que les contributions de A et D, bien que ces derniers aient des coordonnées plus extrêmes. La raison tient dans les effectifs marginaux : B et C ont, certes, des coordonnées plus faibles mais ils ont des poids beaucoup plus élevés.

On voit sur cet exemple l'intérêt d'examiner les contributions en complément des graphiques de l'analyse des correspondances. Tempérons tout de fois ceci en disant que dans notre exemple on a de très grandes différences d'effectifs marginaux qui vont de 1 à 10 et même plus que cela. On en conclura que, en analyse des correspondances, si l'on a des différences d'effectifs marginaux très importantes, la consultation des contributions est indispensable ; en revanche si les effectifs marginaux sont peu différents, la consultation des contributions apporte très peu par rapport aux coordonnées, les coordonnées représentant bien les contributions.

Diapositive 38

Comme en ACP, il est facile de considérer des éléments supplémentaires puisqu'il suffit d'appliquer la formule de transition pour calculer par exemple les coordonnées d'une colonne supplémentaire. On a ajouté une colonne pour la médaille Fields qui est souvent considéré comme l'équivalent d'un prix Nobel en mathématiques.

On peut voir que les mathématiques se positionnent du côté des prix Nobel de la paix et de littérature. Doit-on retrouver ici la proximité entre philosophie et mathématiques ?

On peut voir aussi que les mathématiques se positionnent du côté de la France et de la Russie qui sont deux pays où les mathématiques sont une discipline particulièrement mis en avant.

Diapositive 39

Mentionnons ici une propriété tout à fait remarquable de l'analyse des correspondances. Cette propriété s'appelle l'équivalence distributionnelle. Elle est particulièrement précieuse dans l'analyse de tableaux lexicaux. Si deux lignes (ou deux colonnes) ont exactement le même profil, on peut avoir envie de les regrouper. Quand on regroupe celles-ci, on a donc plus qu'une seule ligne (on a remplacé les 2 lignes par une seule en additionnant les effectifs). On peut alors se poser la question : quel tableau faut-il analyser ? L'analyse des correspondances du premier tableau ou du deuxième tableau. L'équivalence distributionnelle nous dit que ces deux analyses conduisent exactement au même résultat. Ceci est très important lors de l'analyse de tableaux lexicaux, et cela dépassionne tout à fait le débat, "faut-il, ou non, regrouper les mots au singuliers et au pluriels, les conjugaisons des verbes, certains mots synonymes ?" Grâce à l'équivalence distributionnelle, on sait que si ces mots ont le même profil, il revient au même de les regrouper ou non. Et bien entendu, s'ils n'ont pas le même profil, il ne faut pas les regrouper car ils véhiculent des notions différentes. Une vidéo sur l'analyse textuelle est proposée en complément de ces vidéos de cours.

Diapositive 40

Terminons cet exposé par quelques considérations sur le nombre maximum d'axes d'inertie non nulle en analyse des correspondances. Prenons le point de vue du nuage des lignes comportant I points dans un espace à J dimensions. Comme il y a J dimensions, on peut penser dans un premier temps que l'on peut extraire J axes d'inertie non nulle. En fait il n'en est rien car ce que nous analysons ce sont des profils. Les profils ont une particularité, c'est que la somme des coordonnées vaut 1 ; le nuage est donc inclus dans un sous espace de dimension $J-1$. Ainsi, le nombre maximum d'axes d'inertie non nulle est inférieur ou égal à $J-1$. Si on adopte maintenant le point de vue du nombre de points, il nous faut représenter I points. Si on a 2 points, on peut les représenter par un seul axe. De façon plus générale on saura parfaitement représenter à coup sûr I points avec au plus $I-1$ dimensions.

Finalement le nombre maximum d'axes d'inertie non nulle en analyse des correspondances est inférieur au minimum des 2 nombres suivants : le nombre de lignes - 1 et le nombre de colonnes - 1. Ainsi, dans l'exemple sur le prix Nobel, comme nous avons un tableau avec 8 lignes et 6 colonnes, on était sûr avec cinq axes d'avoir un pourcentage d'inertie de 100%.

Quelles conséquences pour le Φ^2 qui est égal à la somme des valeurs propres sachant que ces valeurs propres sont toutes inférieures ou égales à 1. Et bien le Φ^2 est nécessairement inférieur ou égal au nombre maximum d'axes d'inertie non nulle, c'est-à-dire le minimum des deux nombres suivants : $I-1$ et $J-1$.

D'où l'idée de rapporter ce Φ^2 à sa valeur maximum. On obtient ainsi un indicateur, le V de Cramer, qui est compris entre 0 et 1 et qui mesure l'intensité de la liaison entre les deux variables qualitatives. On a donc deux indicateurs d'intensité de liaison, le Φ^2 et le V de Cramer. Ils sont très proches, le V de Cramer a l'avantage de varier entre 0 et 1.

Calculons ce V de Cramer pour les prix Nobel. On trouve 0.03, c'est une valeur faible qui nous indique que l'on est très très loin d'une association exclusive entre modalités. Si l'on calcule maintenant ce même

indicateur le V de Cramer sur les tableaux des 3 saveurs, on trouve des valeurs de 0.69 et 0.52, considérablement plus élevées, ce qui signifie qu'on a une liaison assez étroite entre les deux variables.

Diapositive 41

Quel bilan pouvons-nous tirer de l'exemple sur les prix Nobel ? Même sur des données de petite taille, l'analyse des correspondances apporte une visualisation synthétique de l'écart à l'indépendance qui aide la compréhension du tableau. Ceci sera vrai, a fortiori, avec de grands tableaux. Alors dans ces données, on peut dire que l'essentiel de l'écart à l'indépendance est structuré par une opposition entre les prix scientifiques et les autres, et dans une moindre mesure à une opposition entre la physique/chimie et les sciences économiques. La position des pays illustre leur spécificité par rapport aux prix qu'ils obtiennent : le Royaume-Uni, au centre du graphique, n'est spécialisé dans aucune discipline et obtient des prix dans des proportions comparables à la proportion de prix distribués. France et Italie sont plutôt spécialisés dans les prix Nobel de la paix et de littérature, Japon et Allemagne sont spécialisés en physique/chimie tandis qu'USA et Canada sont spécialisés en sciences économiques.

Diapositive 42

Conclusion sur cet exposé. Pour étudier la liaison entre deux variables qualitatives, on construit un tableau de contingence, ça c'est la base. La liaison réside dans l'écart entre le tableau de contingence et le modèle d'indépendance. L'analyse des correspondances construit un nuage des lignes et un nuage des colonnes dont l'inertie totale mesure l'intensité de l'écart à l'indépendance ; elle décompose cette inertie totale sur une suite de directions d'importance décroissante représentant chacune un aspect synthétique de la liaison entre les deux variables. Enfin l'AFC fournit une représentation simultanée des lignes et des colonnes dans laquelle la position d'un point reflète sa participation à l'écart à l'indépendance.

Diapositive 43

Pour aller plus loin on consultera le livre présentant l'analyse des correspondances dans le même esprit que cette vidéo.

Vous avez vu toutes les vidéos de cours sur l'analyse des correspondances, vous pouvez voir maintenant la vidéo sur comment mettre en œuvre l'analyse des correspondances sous FactoMineR et faire les exercices.