Transcription de l'audio du cours de classification

Première partie. **La classification ascendante hiérarchique**
Diapositives 1 à 13
Pages 2 à 7

Deuxième partie. **Exemple et choix du nombre de classes**
Diapositives 14 à 20
Pages 8 à 10

Troisième partie. **Méthodes de partitionnement et compléments**
Diapositives 21 à 27
Pages 11 à 13

Quatrième partie. **Caractérisation des classes**
Diapositives 28 à 40
Pages 14 à 18

Première partie. La classification ascendante hiérarchique

(Diapositives 1 à 13)

Diapositive 1

Cette semaine nous nous intéressons aux méthodes de classification, la classification ascendante hiérarchique et une méthode de partitionnement, les K-means.

Diapositive 1bis (plan)

L'ensemble des vidéos de cours de cette semaine aborde les points suivants : après une brève introduction sur les données rencontrées en classification et les objectifs de la classification, nous verrons quels sont les principes généraux de la classification, notamment de la classification ascendante hiérarchique. Quels critères peuvent être utilisés et quels algorithmes ? Ensuite nous décrirons une méthode de partitionnement, la méthode très connue des K-means, et enfin nous verrons différents compléments sur l'utilisation conjointe de la classification et des K-means, de la classification sur des données de grandes dimensions ou de la classification sur des données qualitatives. Nous terminerons cet exposé en proposant une méthode pour caractériser les individus d'une même classe.

Diapositive 2 (plan suite)

Commençons dans cette vidéo par donner quelques définitions et par présenter la classification ascendante hiérarchique.

Diapositive 3

La classification c'est l'action de constituer ou de construire des classes, des groupes ou des catégories; et les classes sont des ensembles d'individus ou d'objets qui possèdent des traits de caractères communs. Des traits de caractères communs, c'est-à-dire que ce sont des individus qui se ressemblent du point de vue de l'ensemble des caractères qui les décrivent. Alors des classifications vous en connaissez. Par exemple, le règne animal est un arbre de classification particulier, le disque dur d'un ordinateur, organisé avec des répertoires, des sous-répertoires, des sous-sous-répertoires et finalement des fichiers. Toute cette arborescence est en fait une classification. Des classes vous en connaissez aussi : les classes sociales, les catégories socio-professionnelles, les classes politiques. Donc on va regrouper des individus qui ont des caractéristiques communes. Et donc cela amène à deux types de classification. Des classifications appelées hiérarchiques pour lesquelles on cherchera à construire un arbre hiérarchique pour voir comment s'organise les objets ou les individus. On va parler ici de classification ascendante hiérarchique. Et puis des classifications de type méthode de partitionnement où on va essayer uniquement de constituer des groupes d'individus qui se ressemblent et de constituer une partition.

Diapositive 4

Voici un exemple d'arbre hiérarchique, l'arbre hiérarchique du règne animal. On trouve dans le règne animal : des embranchements. Différents embranchements (les arthropodes, les cnidaires, les

vertébrés); dans les vertébrés, on trouve différents vertébrés : les mammifères, les oiseaux, les reptiles, les amphibiens, les poissons. Parmi les mammifères, on trouve les monotrènes, les édentés, les tubulidentés, les carnivores etc. et si on descend dans la hiérarchie on trouve des classes de plus en plus homogènes. Dans le bas de la hiérarchie, les espèces ont des caractéristiques très communes. Juste au dessus on a des genres qui regroupent plusieurs espèces, dans des classes qui sont plus grosses mais les individus d'une même classe se ressemblent un petit peu moins.

Diapositive 5 (plan)

Décrivons les principes généraux de la classification ascendante hiérarchique. Nous allons voir quels sont les critères utilisés, les algorithmes pour construire un arbre hiérarchique et enfin comment quantifier la qualité d'une partition avant de proposer une méthode pour des données euclidiennes : la méthode de Ward.

Diapositive 6

Quelles sont les données sur lesquelles on va pouvoir construire une classification ascendante hiérarchique ? Ce sont les mêmes données qu'en analyse en composantes principales à savoir des tableaux de données avec des individus en lignes et des variables quantitatives en colonnes. Les variables sont quantitatives mais nous verrons en fin d'exposé comment faire quand les variables sont qualitatives. L'objectif de la classification est de produire une arborescence qui met en évidence les liens hiérarchiques entre les individus ou entre des groupes d'individus. Par exemple sur l'arbre hiérarchique ici à droite, les individus A et C sont très proches et sont aussi assez proches de l'individu B. Ces trois individus forment une classe. D'autre part, les cinq individus D, E, F, G et H se ressemblent et à l'intérieur de cette classe, les individus F et G se ressemblent encore plus. Ces représentations sous forme d'arbre permettent également de détecter un nombre de classes naturel dans une population. Par exemple dans l'arbre proposé on peut décomposer les individus en deux classes, la classe A-B-C et la classe D-E-F-G-H.

Diapositive 7

Alors quels sont les critères de la classification ? Il faut, pour réaliser une classification, définir une mesure de ressemblance entre deux individus. Quand est-ce que deux individus sont très proches ? Quand est-ce qu'on va les mettre dans une même classe ? Une distance très connue et naturelle lorsqu'on visualise des données est la distance euclidienne. Nous avons déjà utilisé cette distance en ACP et nous verrons qu'en classification, sur des tableaux avec des individus en lignes et des variables en colonnes, c'est une distance qui est aussi naturelle. Il existe également des indices de similarités souvent associés à un domaine d'application spécifique. En écologie, l'indice de Jaccard, par exemple, est très utilisé; il existe de nombreuses mesures de ressemblance entre individus. Dans ce cours, on va principalement s'intéresser à la distance euclidienne puisqu'on fera le lien avec les méthodes d'analyse factorielle qui fournissent des représentations, elles aussi, euclidiennes. Donc là on s'est intéressé à la ressemblance entre deux individus. On doit aussi définir la ressemblance entre groupes d'individus.

Sur le petit schéma suivant, on voit deux groupes d'individus et une première mesure de ressemblance : le saut minimum, on parle aussi de lien simple. C'est la distance ici en rouge. La distance minimum entre deux groupes est égale à la plus petite distance entre un élément du

premier groupe et un élément du second. Une autre mesure de distance entre deux groupes est ce qu'on appelle le lien complet : on va prendre cette fois la plus grande distance entre un individu du 1er groupe et un individu du 2ème. On définira dans la suite de l'exposé une autre mesure de ressemblance appelée le critère de Ward. Il existe plusieurs mesures de ressemblance entre individus et plusieurs mesures de ressemblance entre groupes d'individus. Le choix de la mesure de ressemblance modifie la classification que l'on obtient. Donc selon les données, on utilisera certaines distances entre individus et certaines mesures de ressemblance entre groupes d'individus.

Diapositive 8

A partir d'un exemple simple, nous allons construire, à la main, un arbre hiérarchique afin de comprendre le fonctionnement de l'algorithme. Nous considérons ici 8 points, A, B, C, D, E, F, G et H et les coordonnées de ces points sur 2 dimensions. Nous pouvons ainsi représenter les points dans un plan afin de visualiser les distances entre ces 8 points.

Donc dans un premier temps chaque point représente une classe constituée d'un seul individu. Pour cette raison, chaque point est entouré d'une petite ellipse qui ne contient qu'un seul point. La première chose à faire est alors de calculer la distance entre les points. On utilise ici la distance euclidienne. Donc j'ai une matrice de distances : par exemple, entre le point A et le point B, j'ai une distance de 0.5 entre A et C, j'ai une distance de 0.25; entre B et C 0.56, etc.

Première étape : on va chercher la plus petite distance dans cette matrice de distances. La plus petite distance correspond à 0.25 et est celle entre les points A et C.

On commence ainsi à construire l'arbre hiérarchique en regroupant les 2 points A et C. Le regroupement de ces 2 points se fait à une hauteur de 0.25, c'est -à-dire la distance entre ces 2 points. Et donc on a maintenant des groupes d'individus constitués d'un seul individu plus un groupe constitué des individus A et C.

On va alors calculer la distance entre chaque individu et le groupe A-C. Pour ce faire, nous allons utiliser la mesure de ressemblance du saut minimum. La distance entre A-C et l'individu B, si je considère le critère du saut minimum, ça va être la distance de 0.5. En effet la distance entre A et B vaut 0.5 et la distance entre B et C vaut 0.56. Donc la plus petite distance est égale à 0.5. On calcule toutes les distances entre le groupe A-C et chaque individu ce qui donne une nouvelle matrice de distances.

On cherche la plus petite distance de cette nouvelle matrice. C'est la distance entre le groupe A-C et l'individu B.

Donc on va regrouper ces deux groupes d'individus, et les regrouper à une hauteur de 0.5 dans l'arbre. Donc j'ai maintenant un groupe A-B-C et 5 groupes constitués d'un individu.

On calcule la distance entre le groupe A-B-C et chacun des individus pour avoir une nouvelle matrice de distances et dans cette nouvelle matrice de distances, la plus petite distance nous indique qu'il faut regrouper F et G.

On regroupe donc ces 2 points à une hauteur de 0.61.

On continue en calculant une nouvelle matrice de distances qui contient le groupe F-G. La plus petite distance cette fois est entre D et E avec une distance de 1.

On regroupe D et E à une distance de 1. On a alors à cette étape les groupes A-B-C, D-E, F-G et H.

On calcule la nouvelle matrice de distances. La plus petite distance est entre F-G et le point H.

On regroupe donc le groupe F-G et le point H à une hauteur de 1.12.

On continue avec les 3 groupes restants. La plus petite distance est entre le groupe D-E et le groupe F-G-H.

On les regroupe à la hauteur de 1.81.

Il ne reste plus que 2 groupes que l'on regroupe à la distance de 4.07.

Et nous finissons la construction de l'arbre avec ce dernier regroupement. Ainsi itérativement, nous avons construit l'arbre hiérarchique en regroupant pas à pas les deux groupes les plus proches. Evidemment, construire un arbre hiérarchique est ici relativement aisé car on a peu d'individus. Mais dès que le nombre d'individus est plus grand il sera difficile de construire un tel arbre à la main et il sera nécessaire d'utiliser des programmes.

Diapositive 9

Alors une fois qu'on a construit cet arbre, les arbres hiérarchiques, comme tous les arbres finissent par ... être coupés. Ici, on va vouloir couper l'arbre pour constituer des classes.

En définissant un niveau de coupure sur un arbre, on définit une partition. Dans l'arbre suivant, le niveau de coupure, représenté par le trait noir, définit une partition en quatre classes. En définissant le niveau de coupure on définit un nombre de classes.

Alors évidemment, vu le mode de construction de l'arbre, la partition n'est pas nécessairement optimale. En effet on a pris en compte lors de la construction de l'arbre une contrainte de hiérarchie entre individus ou groupes d'individus, ce qui n'est pas utile pour définir une partition. En levant cette contrainte de hiérarchie, il est possible d'améliorer la partition, ce que nous verrons à la fin de cet exposé. Toutefois, si la partition obtenue en coupant un arbre hiérarchique n'est pas nécessairement optimale, c'est souvent une partition qui est de bonne qualité.

Diapositive 10

Une partition va être de bonne qualité si les individus d'une même classe sont très proches, s'ils ont des caractéristiques communes? Une partition est bonne également si les individus de deux classes différentes sont éloignés, ont peu de caractéristiques communes. Alors comment traduire mathématiquement ces deux idées ?

Deux individus d'une même classe sont proches si la variabilité intra-classe est petite. Ca veut dire que, à l'intérieur d'une classe, il y a très peu de variabilité, les individus se ressemblent. Et la deuxième affirmation, les individus de deux classes différentes sont éloignés si d'une classe à l'autre il y a une grande variabilité. Ca veut dire qu'on veut une variabilité inter-classes grande. Donc on a envie d'avoir à la fois une variabilité intra-classe petite et une variabilité inter-classes grande.

Donc ça nous donne deux critères. Alors lequel choisir ? Il est toujours délicat de choisir, mais là, ces deux critères n'en font en fait qu'un.

Diapositive 11

En effet, l'inertie totale, la variabilité totale, représentée en bleu sur le schéma, se décompose en une variabilité intra-classe (représentée en noire) plus une variabilité inter-classes (représentée en rouge). Donc ici, Xiqk est la valeur prise par l'individu i de la classe q pour la variable k; X barre k est la moyenne de la variable k; X barre qk est la moyenne de la variable k dans la classe q, pour les individus de la classe q. L'inertie intra, c'est la variabilité à l'intérieur de la classe et cela correspond à la somme de tous les écarts (au carré) entre les Xiqk et le X barre qk. L'inertie inter c'est la somme des écarts (au carré) entre les moyennes de chaque classe, X barre qk, et les moyennes de chaque variable X barre k. Et donc grâce au théorème de Huygens, on sait que l'inertie totale est égale à l'inertie inter plus l'inertie intra. On peut raisonner variable par variable et considérer la somme sur l'ensemble des variables pour bien comprendre cette équation inertie totale = inertie inter + inertie intra. Par conséquent, minimiser l'inertie intra revient à maximiser l'inertie inter puisque l'inertie totale reste constante.

Donc finalement on a vraiment un seul critère. On peut se focaliser sur l'inertie intra et la minimiser ou sur l'inertie inter et la maximiser.

Diapositive 12

Ceci nous suggère un indicateur de la qualité d'une partition : le ratio inertie inter sur inerte totale. Ce ratio varie entre 0 et 1 et plus il est proche de 1, meilleure est la partition.

L'Inertie inter sur l'inertie totale est égale 0 quand, pour toutes les variables k, les X barre qk sont égaux aux X barre k. Cela signifie que toutes les classes ont la même moyenne, et ce pour chaque variable. Alors évidemment si toutes les classes ont la même moyenne c'est une partition qui ne sépare pas les classes et qui ne permet pas de classifier.

Si l'inertie inter sur l'inertie totale est égale à 1 ça veut dire que l'inertie intra est nulle. Cela signifie que, à l'intérieur d'une classe, les individus sont identiques. Si les individus à l'intérieur d'une classe sont identiques, ça veut dire que les classes sont très homogènes, et ça c'est idéal pour classifier.

Alors attention toutefois : ce critère inertie inter sur inertie totale ne peut pas être jugé en absolu. En effet, ce critère dépend du nombre d'individus et du nombre de classes. Si on augmente le nombre de classes, il est plus facile d'avoir des classes homogènes. Au contraire, si le nombre de classes est petit, à l'intérieur de chaque classe la variabilité sera plus grande. Il faut donc relativiser ce critère par rapport au nombre d'individus et au nombre de classes.

Diapositive 13

Ce critère de qualité d'une partition suggère une nouvelle méthode pour construire une classification ascendante hiérarchique. Cette méthode a été développée par Ward, et s'appelle la méthode de Ward. Le principe est le suivant : on part d'une classification où une classe correspond à un individu. Si une classe correspond à un individu, alors à l'intérieur de la classe il n'y a pas de variabilité intra et l'inertie inter est donc égale à l'inertie totale. La partition est donc idéale. L'objectif est alors de

choisir deux classes a et b telles que leur agrégation minimisent la diminution de l'inertie inter. En effet, l'inertie inter ne peut que diminuer lors d'un regroupement de deux classes. Et on va chercher à minimiser cette diminution de l'inertie inter.

Voyons comment s'écrit la somme de l'inertie d'une classe a et de l'inertie d'une classe b en fonction de l'inertie de l'agrégation de ces 2 classes. L'inertie de a plus l'inertie de b est égale à l'inertie de la réunion de ces deux classes moins une certaine quantité (m_a * m_b)/ (m_a + m_b) multiplié par $d^2(a,b)$. m_a est le nombre d'individus de la classe a, m_b le nombre d'individus de la classe b et $d^2(a,b)$ est la distance entre les centres de gravité de la classe a et de la classe b. Comme on veut que l'inertie de la réunion des classes a et b soit la plus proche possible de l'inertie de a plus l'inertie de b, il suffit donc de minimiser cette dernière quantité. Cette dernière quantité contient deux choses : des poids et une distance au carré.

Cette quantité tout d'abord (m_a * m_b)/ (m_a + m_b) va permettre de regrouper des objets de faible poids et éviter ce qu'on appelle des effets de chaîne. On a ici un petit graphique avec deux classes : la classe bleue et la classe rouge et les arbres hiérarchiques avec le saut minimum à gauche et le critère de Ward à droite. Quand les classes sont bien séparées on retrouve les 2 mêmes classes avec les deux critères. En revanche, avec ces deux mêmes classes plus beaucoup d'individus qui vont de la 1ère classe à la 2ème, et en utilisant le saut minimum, l'arbre hiérarchique met en évidence un effet de chaîne qui conduit à regrouper les individus de proche en proche. Donc l'arbre ne met pas du tout en évidence deux classes dans cette exemple. Avec Ward en revanche, et grâce à cette pondération, les deux classes rouge et bleue restent séparées.

Le 2ème terme de la quantité à minimiser est $d^2(a,b)$. C'est la distance entre les barycentres des classes a et b. Il est tout à fait naturel de regrouper des classes qui ont des centres de gravité qui sont proches. L'intérêt est immédiat ici pour la classification : on regroupe des classes qui sont très proches.

Nous avons vu comment fonctionne la classification ascendante hiérarchique, nous verrons dans les vidéos suivantes, comment la mettre en œuvre sur un exemple, comment utiliser la classification ascendante hiérarchique pour déterminer un nombre de classes et comment construire une partition des individus. Nous verrons également dans la dernière vidéo comment caractériser les individus d'une même classe. N'oubliez pas de faire les quiz pour vous assurer que vous avez bien compris les différentes notions abordées dans cette vidéo.

Deuxième partie. Exemple et choix du nombre de classes (Diapositives 14 à 20)

Nous avons vu la dernière fois comment construire un arbre hiérarchique, nous allons cette fois l'appliquer sur un exemple.

Diapositive 14 (plan)

Alors prenons un exemple pour illustrer la classification. C'est un exemple que nous avons déjà utilisé dans une vidéo sur l'ACP.

Diapositive 15

Voici le tableau de données. On a 15 villes de France en lignes et 12 variables qui correspondent aux températures mensuelles moyennes mesurées sur 30 ans; on a de plus 2 variables, la latitude et la longitude, qui ne vont pas être utilisées pour construire les classes mais qui pourront servir éventuellement lors de la caractérisation des classes. On va donc construire la classification sur la base des données de températures uniquement. En construisant une classification, on va chercher à regrouper des villes qui ont des profils météo similaires et dans une deuxième étape on cherchera à caractériser les différents groupes de villes.

Diapositive 16

Voici l'arbre hiérarchique construit avec la distance euclidienne et le critère de Ward. On voit par exemple que les villes de Rennes et Nantes ont des profils météo très proches. Les températures de ces 2 villes se ressemblent tous les mois de l'année. On voit également un groupe Toulouse - Bordeaux - Nice - Montpellier - Marseille assez homogène. Plus dans le détail, on voit que Montpellier et Marseille dans ce groupe sont les plus proches. On peut voir ainsi les proximités entre villes et entre groupes de villes. Dans le diagramme en haut à droite, on voit l'évolution de l'inertie pour différentes partitions.

Diapositive 17

Examinons ce diagramme de plus près. Ce diagramme montre les pertes d'inerte inter lors d'un regroupement de 2 classes. Plus précisément, il donne la perte d'inertie lors du passage de 15 classes en 14 classes; de 14 classes en 13 classes, etc. et de 2 classes en 1 classe.

Si on somme les pertes d'inertie on trouve la valeur de 12. 12 correspond à la somme des variances des variables du jeu de données car ici on a 12 variables et les variables ont été centrées-réduites. Donc en faisant la somme des pertes d'inertie inter, on retrouve bien l'inertie totale qui est égale à 12. Examinons maintenant l'information apportée par chaque barre de ce diagramme. La 1ère barre ici en bas, la plus grande, donne la perte d'inertie inter lorsqu'on regroupe 2 classes d'individus en 1 seule classe. La perte d'inertie inter est de 7.88, ce qui est très important. Cela veut dire que ce regroupement agrège des individus très différents. On n'a donc pas envie de regrouper ces 2 classes. La barre rouge montre la perte d'inertie lorsqu'on passe de 3 classes en 2 classes. Cette perte d'inerte inter est de 1.56. Cette quantité est relativement importante également et on peut se demander s'il faut ou non faire ce regroupement pour passer de 3 classes à 2 classes. Et le

diagramme montre toutes les pertes d'inertie inter. Les passages de 15 à 14 classes, 14 à 13, etc. du début conduisent à de très faibles pertes d'inertie inter et donc les regroupements sont naturels. La question est alors de savoir jusqu'à quand peut-on regrouper les classes et quand faut-il s'arrêter de regrouper ? Cette question peut simplement se traduire par : combien de classes faut-il faire ?

Diapositive 18

Dans notre exemple, combien doit-on faire de groupes ? Faut-il en faire 2? 3 ? 4 ? On peut se poser la question.

Si on choisit un niveau de coupure ici en orange, on a un découpage en deux groupes.

L'inertie inter, on l'a vu, vaut 7.88 et l'inertie totale 12, ce qui donne un ratio inertie inter sur inertie totale égal à 66 %. C'est-à-dire qu'en séparant les villes en 2 groupes, les villes en rouge, Toulouse, Bordeaux, Nice, Montpellier, Marseille et les villes en bleu Brest, Rennes, Nantes, Grenoble etc. jusqu'à Strasbourg, avec cette séparation des villes en 2 groupes on récupère 66% de l'information qui est contenue dans le tableau de données. Ce découpage en 2 classes résume assez grossièrement les ressemblances entre individus du tableau de données.

Alors à quoi comparer ce pourcentage de 66 % ?

Diapositive 19

En fait, on peut comparer ces 66 % au pourcentage de variabilité expliquée par le premier axe de l'ACP. Le premier axe de l'ACP récupère environ 80 % de l'information du jeu de données. Avec la classification, en séparant juste les villes rouges des villes bleues, on récupère 66 % d'information donc un peu moins d'information qu'avec le premier axe de l'ACP. En effet, le premier axe de l'ACP donne une information plus fine : il sépare Nice de Toulouse ou Bordeaux en donnant une coordonnée plus extrême à Nice. De même, Lille est plus extrême que Nantes. Avec la classification, ces deux villes sont dans la même classe, et avec les regroupements on récupère 66% d'information. Donc on a un résumé plus grossier avec la classification qu'avec l'ACP.

Diapositive 19bis

Alors maintenant, si on sépare les villes froides Brest, Rennes, Nantes, Grenoble, Lyon jusqu'à Strasbourg en 2 groupes, on est en train de faire une classification en 3 groupes. Le passage en 3 classes permet de ne pas perdre l'inertie perdue lors du passage de 2 classes à 3 classes. On récupère une inertie inter de 1.56 soit 13% de l'information. Cette séparation des villes froides en 2 groupes permet de récupérer 13% d'information supplémentaire.

Diapositive 20

Et ces 13%, je peux les comparer au deuxième axe de l'ACP ici puisque effectivement le deuxième axe de l'ACP sépare les villes froides en 2 groupes : les villes en bleu et les villes en vert. Donc avec une classification en trois classes on récupère 66 % + 13 % c'est-à-dire 79% de la variabilité des données. En constituant 3 classes on récupère 79% de la variabilité totale. Avec l'ACP et deux axes on récupère environ 99 % de l'information.

Diapositive 20bis

Le choix du nombre de classes est important car construire une partition avec trop peu de classes risque de conduire à des classes qui ne sont pas homogènes, et au contraire, construire une partition avec trop de classes risque de conduire à des classes qui ne se différencient pas suffisamment. Se pose alors la question cruciale du choix du nombre de classes. On peut déterminer un nombre de classes à partir de l'arbre : on aura tendance à couper l'arbre là où les branches sont assez longues. On peut aussi s'appuyer sur le diagramme des indices de niveau et choisir d'arrêter de regrouper des classes quand le saut est faible, ce qui indique qu'on récupère assez peu d'informations et qu'il n'est plus vraiment utile de regrouper certaines classes. Le nombre de classes dépend également de l'enquête et du nombre d'individus. Si on a une enquête avec des milliers d'individus, il sera intéressant de constituer 5, 6 voire une dizaine de classes. Si maintenant on a beaucoup moins d'individus, on préfère avoir moins de classes. De même, si on classe des individus statistiques très différents, on va préférer constituer plus de classes. Un critère ultime lorsqu'on construit des classes, est l'interprétabilité des classes. Peut-on comprendre et décrire les classes ? Il est inutile de séparer une classe en 2 si on ne peut pas comprendre les 2 classes qu'on est en train de constituer.

Dans cet exemple, nous avons vu que la classification ascendante hiérarchique suggère un nombre de classes. Nous verrons dans la vidéo suivante comment construire une méthode de partitionnement quand le nombre de classes est déterminé.

Troisième partie. Méthodes de partitionnement et compléments (Diapositives 21 à 28)

Diapositive 21 (plan)

On a vu ici comment construire une classification ascendante hiérarchique, un arbre hiérarchique, on va voir maintenant une méthode de partitionnement directe, qui ne nécessite pas la construction d'un arbre hiérarchique. Cette méthode de partitionnement est appelée méthode des K-means ou méthode d'agrégation autour des centres mobiles.

Diapositive 22

Présentons l'algorithme des K-means sur un exemple en 2 dimensions. L'algorithme des K-means nécessite de connaître le nombre de classes Q que l'on souhaite construire. Ici nous allons prendre Q=3 et construire donc une partition en 3 classes.

Au départ ont choisi Q centres de classe au hasard : Rennes, Vichy et Clermont sont les 3 centres de classes choisis au hasard. C'est la première étape, l'étape d'initialisation de l'algorithme.

Ensuite j'affecte tous les individus, tous les points, au centre de classe le plus proche. Donc tous les individus ici en rouge sont plus proches de Rennes que de Vichy ou Clermont. Tous les individus en vert sont plus proches de Vichy que de Clermont et Rennes, et tous les individus en noir sont plus proches de Clermont. On a ainsi réparti les individus dans 3 classes;

On calcule maintenant les centres de gravité de chacune de ces classes. Voici les centres de gravité des classes; ces centres de gravité ne sont plus des individus mais ce sont des points fictifs. Et on va itérer ces deux étapes d'affectation des points au centre le plus proche puis de calcul des centres de gravité des classes.

On va affecter à nouveau les points au centre le plus proche. Calculer les centres de gravité des nouvelles classes. Et on continue d'itérer. On affecte les villes aux centres de gravité le plus proche. On calcule les centres de gravité, les centres de gravité se déplacent. On affecte les points au centre le plus proche.

On calcule les nouveaux centres de gravité ... et on voit que les classes n'ont pas changé et que les centres de gravité ne bougent plus. L'algorithme a alors convergé. On a ainsi constitué les 3 classes. Voici le fonctionnement de l'algorithme des K-means. C'est un algorithme très rapide, qui peut être lancé avec beaucoup d'individus mais qui a 2 défauts. Le premier est qu'il faut connaître le nombre de classes a priori. Le second est que la partition dépend de l'initialisation et du choix des centres de classes au hasard. D'une initialisation à l'autre, la partition peut être très différente. Pour remédier à ce problème, on lancera plusieurs fois l'algorithme avec des initialisations différentes et on conservera la meilleure partition.

Diapositive 23 (plan)

Nous avons vu comment construire une classification hiérarchique et comment construire une partition sur les individus. Dans cette partie, nous allons voir quelques compléments sur la classification, et notamment comment consolider une partition, comment construire une partition

quand on a un jeu de données en grandes dimensions, comment faire quand les variables sont qualitatives et enfin quel est l'intérêt de combiner analyse factorielle et classification.

Diapositive 24

On peut consolider la partition obtenue par la classification ascendante hiérarchique. En effet nous avons vu que la partition obtenue en coupant un arbre hiérarchique n'est pas optimale parce que l'on prenait en compte toute une hiérarchie entre les individus et groupe d'individus. On peut alors consolider la partition en utilisant les K-means. On part de la partition obtenue par le découpage de l'arbre hiérarchique comme initialisation dans l'algorithme des K-means et on va itérer quelques étapes de K-means jusqu'à convergence. Cela améliore nécessairement la partition puisque, par construction, à chaque étape des K-means le critère diminue. L'inconvénient par contre est que les liens hiérarchiques entre individus sont perdus : en effet, certains individus vont changer de classe par rapport à la partition obtenue depuis l'arbre et donc on perd complètement la hiérarchie.

Diapositive 25

La construction de l'arbre hiérarchique nécessite de calculer à chaque étape de nombreuses distances entre individus et groupes d'individus, et pour des données en grandes dimensions, cet algorithme est trop long et fonctionne mal. Alors comment faire donc pour gérer des données de grandes dimensions ? Si le nombre de variables est grand, on peut dans un premier temps faire une ACP et ne conserver que la première dimension de l'ACP. L'ACP va permettre de concentrer l'information sur les premières dimensions factorielles. Et par suite, on va pouvoir faire la classification sur un tableau avec des individus et quelques dimensions factorielles en colonnes. On se ramène alors à un tableau individu x variables quantitatives de taille raisonnable.

Si maintenant on a beaucoup d'individus, l'algorithme de classification est trop long et en pratique on ne peut pas construire l'arbre hiérarchique. Une possibilité est alors de travailler par étapes : d'abord on construit une partition très grossière avec la méthode des K-means en construisant une centaine de classes par exemple. Cette première étape, du fait du grand nombre de classes, assure le regroupement d'individus globalement proches. Il est alors possible de considérer les centres des classes et de construire une classification hiérarchique à partir de ces centres de classe. Evidemment, on utilisera les effectifs des classes dans le calcul de l'arbre. L'arbre obtenu correspond au haut de l'arbre hiérarchique qui serait obtenu à partir de tous les individus, le bas de l'arbre hiérarchique étant perdu. Cependant, avec 1 million d'individus, le bas de l'arbre n'est pas lisible et est inutile car seul le haut de l'arbre est en général commenté.

Voici un exemple avec 300 individus. A gauche, on a l'arbre construit avec tous les individus et à droite l'arbre construit après avoir préalablement construit une partition en 50 classes. On voit que les hauts des 2 arbres sont très similaires et seront commentés à peu près de la même façon.

Diapositive 26

Nous avons jusqu'à maintenant traité le cas où les variables sont quantitatives. Si maintenant les variables sont qualitatives, comment peut-on faire ? Il y a 2 stratégies. La première stratégie est de se ramener à des variables quantitatives en utilisant l'analyse des correspondances multiples. L'ACM va permettre de passer d'un tableau de variables qualitatives à un tableau de coordonnées des individus sur les dimensions et les coordonnées des individus ce sont bien des variables quantitatives. On peut

alors conserver les premières dimensions de l'ACM, ou éventuellement toutes les dimensions. Une fois qu'on s'est ramené à un tableau individus x variables quantitatives, on peut réaliser la classification comme précédemment.

Une autre stratégie consiste à utiliser des mesures adaptées spécifiquement aux données qualitatives. De nombreux indices existent, indices de similarité ou de dissimilarités, comme par exemple l'indice de Jaccard. Ces indices vont permettre de calculer une matrice de distances ou de dissimilarités entre individus. Et à partir de cette matrice de distances, on peut construire un arbre hiérarchique.

Diapositive 27

Nous venons de voir que, si les données sont qualitatives, l'ACM renvoie des composantes principales qui sont quantitatives et sur lesquelles on va pouvoir faire directement une classification pour utiliser le même algorithme de classification. Mais les méthodes d'analyses factorielles ont d'autres avantages. Soulignons ici l'intérêt d'un enchaînement analyse factorielle - classification. Pourquoi il est intéressant de faire une analyse factorielle avant de faire une classification ?

L'analyse factorielle va concentrer l'information sur les premières composantes principales, les dernières composantes n'étant que du bruit, c'est-à-dire de l'aléatoire. Et donc il est possible d'éliminer ces dernières composantes avant de faire la classification. Ainsi, on va éliminer l'aléatoire, le bruit, avant de faire la classification, ce qui permettra d'obtenir une classification qui, en pratique, est plus stable (plus stable dans le sens où en enlevant ou ajoutant quelques individus, les classes restent les mêmes).

Et autre intérêt des méthodes factorielles utilisées en complément de la classification : la possibilité de représenter l'arbre hiérarchique ainsi que les classes sur le plan factoriel. Si on a sur un même graphe, les points sur le plan factoriel, l'arbre hiérarchique et les individus coloriés en fonction de leur appartenance à différentes classes, on a trois informations : on a notamment une vision continue grâce au plan de l'analyse factorielle, on a une vision discontinue avec la classification ascendante hiérarchique qui nous permet également de voir ce qui se passe sur les dimensions 3 - 4 de l'analyse factorielle par exemple. Dans notre exemple, les individus sont parfaitement représentés sur le plan 1-2 et donc les proximités entre individus sur le plan 1-2 sont les proximités des individus dans l'espace. Mais il se peut que certains individus semblent proches sur le plan factoriel et soient assez éloignés sur une troisième et une quatrième dimension ce que la classification pourra mettre en évidence. Cela pourrait arriver dans d'autres exemples où 3 ou 4 dimensions sont nécessaires pour interpréter les résultats de l'ACP. On a alors une vision très synthétique avec les classes et un peu plus fine avec le plan factoriel.

Nous avons vu dans cette vidéo comment construire des classes, nous verrons dans la vidéo suivante comment caractériser les individus d'une même classe.

Quatrième partie : Caractérisation des classes

(Diapositives 28 à 40)

Diapositive 28 (plan)

Maintenant que nous avons vu comment constituer des classes, quels étaient les individus de chaque classe, intéressons-nous à la caractérisation de classes d'individus.

Diapositive 29

On peut tout d'abord donner, le parangon de chaque classe, c'est-à-dire l'individu le plus proche du centre de la classe. Pourquoi utiliser l'individu le plus proche du centre la classe ? Parce que le centre de gravité d'une classe est un individu moyen fictif et il est préférable d'utiliser un vrai individu pour comprendre comment se comporte la moyenne d'une classe. Donc par exemple ici, on voit que, dans la classe 1, c'est Montpellier qui est le plus proche du centre de la classe. Il est à une distance de 0.42 du centre de gravité de la classe, et donc si on veut interpréter comment se comporte la moyenne de la classe on peut regarder l'individu Montpellier. Pour la classe 2, c'est Rennes qui est le plus proche du centre de la classe. Et pour la classe, 3 c'est Vichy. Les parangons sont assez utiles pour ceux qui connaissent bien les données.

On peut représenter ces parangons sur le plan factoriel. On a ici le centre de gravité de la classe 2 et le parangon de cette classe qui est Rennes.

Diapositive 30

Cette caractérisation des classes par l'individu le plus proche du centre de la classe ne suffit pas. Nous aimerions aussi caractériser les classes par les variables. L'objectif est alors de trouver les variables qui caractérisent le mieux la partition dans son ensemble ou qui caractérise le mieux une classe d'individus. Donc les questions sont : quelles sont les variables qui caractérisent le mieux la partition ? Comment caractériser les individus de la classe 1 ? Quelles variables les caractérisent le mieux ?

Diapositive 31

La partition, c'est-à-dire l'appartenance des individus aux différentes classes, peut être considérée comme une variable qualitative à autant de modalités qu'il y a de classes. Donc chercher les variables qui caractérisent le mieux une partition revient à chercher les variables qui caractérisent le mieux une variable qualitative. Pour chaque variable quantitative, on va construire un modèle d'analyse de variance entre la variable quantitative, qui aura le rôle de la variable réponse, en fonction de la variable de classe, qui aura le rôle de la variable explicative. Donc on fait une analyse de variance de la variable quantitative en fonction de la variable de classe et on construit un test de Fisher pour voir s'il y a un effet de la variable de classe sur la variable quantitative. On peut conserver les variables ayant une probabilité critique inférieure à 5% et trier ces variables par probabilité croissante. On voit dans l'exemple que la variable qui caractérise le mieux les classes est la variable octobre qui a la probabilité critique la plus petite et qui permet donc de bien séparer les classes. Notons que, comme pour la description des dimensions en analyse factorielle, il faut utiliser ces tests avec prudence

puisque la variable octobre a joué un rôle actif dans la construction des classes puisqu'elle a participé au calcul des distances entre individus. Seules les probabilités critiques associées aux variables supplémentaires latitude et longitude peuvent être réellement interpréter comme usuellement. Les probabilités critiques des variables actives restent utiles même si elles ne peuvent pas être utilisées comme des tests classiques. Le critère du test de Fisher permet de trouver les variables qui caractérisent globalement la partition. Maintenant on veut comprendre chacune des classes, comment caractériser chaque classe par les variables ?

Diapositive 32

Cette représentation permet de visualiser les données : chaque ligne correspond à une variable, chaque point correspond à une ville. Les points en vert représentent les villes Lyon, Paris, Grenoble, Clermont, Vichy, Strasbourg, Lille; en bleu les villes de Nantes, Rennes, Brest et en rouge les villes de Nice, Marseille, Montpellier, Toulouse, Bordeaux. Et on cherche à voir s'il y a des variables qui permettent de bien caractériser une classe ? Donc est-ce que les valeurs d'une variable sont extrêmes, particulières pour les individus d'une classe ? On a l'impression, par exemple, que, pour la variable janvier, les villes en vert, ont des valeurs plus faibles que toutes les autres, et donc que la variable janvier caractérise la classe verte. Evidemment, il n'est pas question de passer chaque classe et chaque variable une par une à partir d'un tel graphe. On a besoin de procédure automatique.

Diapositive 33

L'idée, pour savoir si une variable caractérise une classe, est la suivante : si les valeurs d'une variable quantitative X pour les individus de la classe q semblent tirés complètement au hasard parmi toutes les valeurs de X, alors la variable X ne caractérise pas la classe car les individus de la classe q ne prennent pas de valeur particulière sur X. Au contraire si les valeurs de X semblent très particulières pour les individus de la classe q, alors l'hypothèse d'un tirage au hasard est douteuse et on dira que la variable X caractérise la classe. Ici sont représentées les valeurs de 2 variables : une variable avec des données au hasard et la variable du mois d'août. On voit que, pour la variable au hasard on a des points bleus, verts et rouges qui sont répartis un peu n'importe comment alors qu'on a l'impression que les points sont plus structurés pour le mois d'août. Le mois d'août semble caractériser la classe rouge par exemple qui a des points assez extrêmes. L'idée est assez intuitive, essayons de construire un test.

Diapositive 34

L'idée générale est de se comparer à un tirage au hasard de nq valeurs parmi N. On a nq valeurs dans la classe q, est-ce que ces valeurs sont choisies au hasard parmi les N valeurs de la population ? Ou alors est-ce qu'on remet en question le fait que ces valeurs soient choisies au hasard dans la population ? Pour le savoir, il nous faut déterminer quelles valeurs peut prendre X barre q, la moyenne d'une classe ? Autrement dit, il nous faut déterminer la loi de X barre q ?

L'espérance de X barre q sous l'hypothèse d'un tirage au hasard c'est la moyenne de toute la population, c'est-à-dire X barre. Pour calculer la variance de X barre q, on doit considérer qu'on tire (sans remise) nq valeurs dans une population de taille finie N. C'est donc l'écart-type de la population, ici, s^2, divisé par le nombre d'individus tirés nq mais comme on est dans une population de taille finie, il y a le terme correctif racine de (N-nq) / (N-1). Quant à la loi de X barre q, comme X

barre q est une moyenne, on peut considérer que sa loi est Normale grâce au théorème central limite.

On peut alors calculer ce qu'on appelle une valeur centrée-réduite X barre q mois X barre sur l'écart type de X barre q, c'est-à-dire la racine carrée de la variance de X barre q. Cette quantité, sous l'hypothèse d'un tirage au hasard, va suivre une loi normale centrée-réduite. Si la valeur-test est comprise entre -1.96 et 1.96, elle peut provenir d'une loi normale centrée-réduite et donc elle n'est pas particulière.

Si maintenant la valeur absolue de la valeur-test est supérieure à 1.96 alors cette valeur-test est particulière, et il est peu probable qu'elle provienne d'une loi normale centrée-réduite. On remettra alors en cause l'hypothèse d'un tirage au hasard des nq valeurs parmi les N, c'est-à-dire que les valeurs de X pour la classe q semblent particulières, et donc ça veut dire que la variable X caractérise la classe q. Et plus l'hypothèse d'un tirage au hasard est douteuse, plus la valeur-test sera grande et plus la variable X caractérisera la classe.

Il est alors possible de trier les variables par valeur-test décroissante.

Diapositive 35

On a les résultats suivants pour la classe 1. La valeur-test la plus grande est pour la variable septembre, ce qui signifie que la variable septembre est la variable qui caractérise le mieux la classe 1. La moyenne des individus pour septembre est de 19.30° alors que la moyenne générale pour tous les individus, y compris ceux de la classe 1, est égale à 17°. On a ensuite l'écart-type dans la catégorie 0.755 et l'écart-type de la population totale 1.79. La dernière colonne donne la probabilité critique associée au test que la valeur-test suit une loi normale. Cette première classe est caractérisée par beaucoup de variables. Toutes les valeurs-tests sont positives, ce qui signifie que les valeurs prises sont plus grandes pour les individus de la classe 1 que pour les individus en général. Cette classe correspond aux villes où il fait chaud tous les mois de l'année.

Diapositive 36

Pour la classe 2, les valeurs-tests sont négatives, la valeur-test la plus extrême est donc la dernière, à savoir celle de l'amplitude thermique annuelle. La valeur test étant négative, la moyenne de la catégorie 12.40 est inférieure à la moyenne générale qui est de 15.90.

Et pour la classe 3, là encore on a les valeurs-tests négatives et il faut lire le tableau en partant du bas. La variable qui caractérise le plus la classe 3 est la variable de janvier qui a la valeur-test la plus extrême, -3.36 Et effectivement, la moyenne dans la classe est de 2.11° contre 3.97° pour la moyenne générale.

Diapositive 37

Peut-on caractériser les classes par des variables qualitatives ? Pour ce faire, il faut à nouveau considérer que la partition est une variable qualitative et on cherche alors le lien entre 2 variables qualitatives. On peut alors construire pour chaque variable qualitative, un test du chi² entre une variable qualitative et la variable de classe. On peut ensuite trier les variables qualitatives par

probabilité critique croissante. Dans l'exemple, il n'y avait qu'une seule variable qualitative, la variable région. Cette variable est liée à la partition des individus.

Diapositive 38

Peut-on maintenant caractériser chaque classe en fonction des modalités ? Prenons l'exemple de la classe 3 et voyons si la modalité Nord-est caractérise cette classe. L'idée est de comparer la proportion de Nord-est dans la classe 3 par rapport à la proportion de Nord-est dans la population globale. On peut donc construire un tableau réduit avec juste la modalité Nord-est et toutes les autres modalités regroupées en une seule modalité qu'on appellera "autre"; de même on a la classe 3 et toutes les autres classes qui sont regroupées.

On peut construire un test pour comparer la proportion de Nord-est dans la classe 3 et dans la population globale. L'hypothèse H0 est de considérer que ces 2 proportions sont égales, contre l'hypothèse alternative la modalité Nord-est est sur-représentée dans la classe ou sous-représentée dans la classe.

Sous l'hypothèse nulle H0, la variable aléatoire notée N_mc qui représente le nombre d'individus de la modalité m dans la classe c, suit une loi hypergéométrique de paramètres n, n_m/n et n_c.

On peut alors calculer la probabilité d'avoir une valeur encore plus extrême que celle observée, c'est-à-dire ici 3.

On obtient alors le tableau suivant : en ligne on a les modalités. Ici nous avons juste une seule ligne correspondant à la modalité Nord-est. La première colonne du tableau donne la proportion de villes appartenant à la classe 3 parmi les villes du Nord-est. Le calcul est donc n_mc sur n_m soit ici 3/3 = 100%. La 2ème colonne donne le pourcentage de villes de la classe 3 qui prennent la modalité Nord-est : ici 3 villes parmi les 7 de la classe 3 sont du Nord-est, soit 42.86%. La colonne global donne la proportion de villes du Nord-est dans la population globale, soit ici n_m/n = 3/15, soit 20%. La 4ème colonne donne la probabilité critique du test, nous indiquant si la modalité caractérise significativement ou non la classe. Ici la probabilité critique est de 0.07 ce qui est légèrement supérieur à 5%, donc on aura tendance à accepter l'hypothèse H0 et donc à considérer que la modalité Nord-est n'est pas caractéristique de la classe, elle n'est ni sous-représentée, ni sur-représentée dans la classe 3. On s'attachera donc à commenter la probabilité critique et à comparer les proportions dans la classe et dans la population pour voir si la modalité est sur ou sous-représentée. La valeur-test, qui est ici juste une conversion de la probabilité critique en quantile de loi normale, donne ces 2 informations. Si sa valeur absolue est supérieure à 1.96 alors la modalité caractérise la classe et si le signe est négatif la modalité est sous-représentée, si le signe est positif la modalité est sur-représentée.

Finalement, comme pour les variables quantitatives, toutes les modalités des variables qualitatives caractérisantes peuvent être triées par probabilité critique croissante.

Diapositive 39

Il est également possible de caractériser les classes par les axes factoriels. Ces axes factoriels sont des variables quantitatives et on peut donc utiliser exactement la même méthodologie que celle que nous venons d'utiliser pour les variables quantitatives. On voit que la classe 1 est caractérisée par le

premier axe factoriel. La valeur test étant positive, les individus de cette classe ont des coordonnées particulièrement extrêmes et positives sur la première dimension factorielle. La moyenne dans la classe est de 3.97 alors que la moyenne générale est de 0. Pour les dimensions factorielles, la moyenne de toutes les coordonnées est toujours égale à 0, puisque les cordonnées factorielles sont centrées. Les individus de la classe 2 prennent des valeurs plus extrêmes sur la dimension 2 et les individus de la classe 3 ont des valeurs plus extrêmes sur les dimensions 1 et 2 et prennent des valeurs négatives sur ces 2 dimensions.

Diapositive 40

Pour conclure, rappelons tout ce que nous avons vu. Les méthodes de classification s'appliquent sur des tableaux individus x variables quantitatives. Mais si les variables sont qualitatives, on peut utiliser l'ACM pour transformer les variables qualitatives en dimensions factorielles qui sont des variables quantitatives.

La classification ascendante hiérarchique fournit un arbre hiérarchique qui montre les distances entre individus et entre groupes d'individus et qui donne également une idée du nombre de classes dans le jeu de données.

On peut alors utiliser une méthode de partitionnement pour consolider les classes obtenues en coupant l'arbre hiérarchique. Les classes sont alors plus stables mais on perd la hiérarchie entre les classes. Ces méthodes de partitionnement peuvent aussi être utilisées comme pré-traitement quand les données sont de grandes dimensions.

Enfin, pour terminer, on peut caractériser les classes par des variables quantitatives et des variables qualitatives. Ces variables peuvent avoir servi à calculer les distances entre individus ou non; en d'autres termes, elles peuvent être actives ou illustratives.

Vous avez vu toutes les vidéos sur la classification, vous pouvez maintenant voir la vidéo sur FactoMineR pour voir comment mettre en œuvre la classification sous FactoMineR et comment caractériser des classes d'individus avec FactoMineR. N'oubliez pas non plus de faire les quiz et les exercices.