Transcription de l'audio du cours d'Analyse en Composantes Principales

Première partie. Données - problématique

(Diapositives 1 à 9)

Diapositive 1 :

Cette semaine, nous vous proposons 3 vidéos de cours qui présentent les principales caractéristiques de l'analyse en composantes principales.

Diapositive 1 bis :

Le plan de l'exposé est le suivant : nous commencerons par décrire les données sur lesquelles on travaille en ACP et dégagerons des objectifs et une problématique.

Ensuite nous verrons que les individus, comme les variables, peuvent être représentés par un nuage de points multidimensionnel et nous verrons comment visualiser ces nuages de points. Enfin nous détaillerons plusieurs aides à l'interprétation très utiles pour analyser les résultats.

Diapositive 2 (plan) :

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

Diapositive 3 :

L'ACP, l'analyse en composantes principales, s'intéresse à des tableaux de données rectangulaires avec en lignes des individus et en colonnes des variables qui sont de nature quantitative. Donc on peut considérer qu'on a I individus et K variables. On définit la moyenne d'une variable par $\overline{X}_k$ qui est la moyenne des X_{ik} et on définit l'écart type d'une variable, s_k, par la racine de *1/I* somme sur i des $(X_{ik} - \overline{X}_k)^2$. On voit ici qu'il y a *1/I* car on estime l'écart-type de la variable pour ces données et ces données seulement, sans chercher à estimer l'écart-type d'une population dont ces données serait un échantillon.

Alors des tableaux de données de ce type, avec des individus en lignes et des variables en colonnes, on en trouve dans de très nombreux domaines d'application.

Diapositive 4 :

Voici une liste d'exemple de ce type de tableaux issus de divers domaines d'application.

Le premier exemple provient de l'analyse sensorielle. Des produits sont décrits par différentes variables qu'on appelle souvent des descripteurs sensoriels comme par exemple l'acidité, l'amertume, la saveur sucrée, etc. Et donc on a un tableau avec différents produits, par exemple des vins, et pour chaque vin on a sa note d'acidité, sa note d'amertume, sa note de saveur sucrée. L'objectif de l'ACP est d'étudier ce tableau afin d'obtenir ce que l'on appelle un espace produit, c'est-à-dire une description sensorielle multidimensionnelle des vins.

En écologie, on a un exemple avec différentes rivières et les variables sont différents polluants donc on a un tableau qui croise des rivières en lignes des polluants en colonnes.

En économie, on peut avoir des années et on suit différents indicateurs économiques d'une année à l'autre, les individus statistiques peuvent être les années mais ce pourrait être des pays.

En génétique, très souvent, on a des patients et les variables qui décrivent les patients sont des gènes. Dans ce cas, le tableau de données contient de très nombreux gènes.

En marketing, on a différentes marques et des indices de satisfaction.

En sociologie on peut avoir des enquêtes budget-temps où les individus statistiques sont par exemple des regroupements d'individus de différentes CSP et les variables sont les différentes activités. Et la mesure est le temps moyen passé par des individus d'une CSP pour une activité donnée.

Donc des tableaux de ce type qui croisent des individus en lignes et des variables en colonnes sont nombreux et issus de domaines très variés.

Diapositive 5 :

Pour illustrer ce cours, nous allons prendre un exemple sur des données météorologiques. En lignes, les individus statistiques sont différentes villes de France, 15 villes de France, et en colonnes les variables quantitatives sont des températures mensuelles moyennes. Ces températures mensuelles moyennes ont été calculées sur 30 ans. Donc, par exemple, à Bordeaux, en janvier il fait en moyenne 5.6°C. Cette valeur de 5.6°C est la moyenne sur tous les jours de janvier pendant 30 ans. On a ainsi 12 variables correspondant au 12 mois de l'année ; on a ensuite 2 variables de positionnement, la latitude et la longitude des villes. L'objectif de notre ACP va être d'étudier ce tableau.

Comment étudier ce tableau ?

Diapositive 6 :

On peut l'étudier de différentes façons : on peut considérer ce tableau comme un ensemble de lignes et chercher les différences et les ressemblances qu'il peut y avoir d'une ligne à l'autre ; ou bien on peut considérer ce tableau comme un ensemble de colonnes et chercher à voir les ressemblances entre colonnes. Pour l'étude sur les individus statistiques, sur les lignes, nous devons définir quand 2 individus se ressemblent et quand est-ce qu'ils se ressemblent du point de vue de l'ensemble des colonnes. Si nous avons beaucoup d'individus, nous voudrons faire un bilan des ressemblances : ces individus se ressemblent, sont très proches du point de vue de l'ensemble des colonnes, ces autres individus sont proches entre eux et différents du premier groupe d'individus. Nous cherchons ici à faire une typologie, une partition des individus, c'est-à-dire à construire des groupes d'individus homogènes du point de vue de l'ensemble des variables. A l'intérieur d'un groupe, les individus se ressemblent et d'un groupe à l'autre ils sont différents.

Diapositive 7 :

Au niveau des variables, de façon symétrique, nous cherchons les ressemblances entre variables. Quelles sont les variables qui apportent une information à peu près identique ? Quelles sont les variables qui apportent des informations différentes? Alors entre variables, plutôt que de ressemblance, on parle souvent de liaison. Et les liaisons les plus connues sont les liaisons linéaires. Ce sont des liaisons simples, très fréquentes, et finalement qui peuvent résumer de nombreuses

liaisons. Même quand une liaison est non linéaire entre deux variables il est fréquent que sur certains domaines la liaison puisse être proche d'une liaison linéaire.

Comment mesurer la ressemblance entre 2 variables ? On peut la mesurer par un indicateur comme le coefficient de corrélation. Deux variables qui ont un coefficient de corrélation très élevé proche de +1 seront des variables qui apportent la même information. Notre objectif dans l'analyse sera de faire un bilan des ressemblances entre variables et de visualiser la matrice de corrélations mais aussi de trouver des indicateurs qui résument beaucoup de variables. Autrement dit, à partir des nombreuses variables du tableau, nous voulons avoir une vision de l'ensemble des liaisons sans passer en revue chaque couple de variables. Cette vision peut se faire par l'intermédiaire de variables synthétiques. Des indicateurs synthétiques, il en existe a priori par exemple la moyenne est un indicateur a priori. Mais ici on va plutôt rechercher des indicateurs a posteriori, c'est-à-dire qu'on va utiliser l'information qui est contenue dans les données pour trouver un indicateur.

Diapositive 8 :

Alors évidemment, comme c'est le même tableau qui est vu en lignes puis en colonnes, il y a un lien entre ces deux études, entre l'étude sur les individus et l'étude des variables.

Dans l'étude sur les individus, nous construisons des groupes d'individus et nous allons chercher à caractériser les différents groupes d'individus, les différentes classes d'individus. Pour caractériser ces groupes, plutôt que de lister tous les individus du groupe, nous préférons utiliser les variables. Sur l'exemple d'analyse sensorielle nous pourrions par exemple dire que ces vins qui se ressemblent sont les vins à la fois acides et amers tandis que ces autres vins sont sucrés et peu acides. Donc, pour cela, on voit bien qu'on a besoin d'une procédure automatique, surtout si on a beaucoup de variables.

De même quand nous étudions les liaisons entre variables nous allons dire que ces différentes variables sont très liées entre elles. Mais ce langage est plutôt abstrait. Et on pourrait l'illustrer par une opposition entre des individus spécifiques, c'est-à-dire des individus qui sont très particuliers qui sont extrêmes. Par exemple, la variable taille et la variable poids sont deux variables très liées. On peut dire qu'il existe une corrélation linéaire forte entre ces variables. Mais on peut aussi illustrer cette liaison en opposant deux individus extrêmes, et en disant : les individus qui sont petits sont légers et les individus qui sont grands sont lourds. L'illustration de la liaison par des individus extrêmes n'est ici pas décisive mais si nous avons beaucoup de variables et que les variables sont moins bien connues, l'illustration par des individus extrêmes est très utile.

Pour résumer, l'ACP est une méthode de statistique exploratoire, de statistique descriptive multidimensionnelle. Cette méthode va synthétiser, résumer, hiérarchiser l'information contenue dans un tableau de données individus x variables quantitatives et pour ce faire elle va fournir une visualisation du tableau de données par des graphiques simples.

Diapositive 9 :

Revenons plus en détail sur les deux façons de considérer le tableau de données. Si je le considère comme un ensemble de lignes, je suis en train d'étudier les individus; si je le considère comme un ensemble de colonnes je suis en train d'étudier les variables. Etudier les individus va revenir à considérer un nuage de points, un point correspondant à un individu. Ce nuage de points évolue dans

un espace de dimension élevée. Si le tableau contient K variables, le nuage de points vit dans un espace à K dimensions.

Dans l'étude sur les variables, une variable est un point dans un espace cette fois à I dimensions (il y a autant de dimensions que d'individus). Chaque variable a donc I coordonnées et correspond donc à un point dans un espace à I dimensions. L'ensemble de ces points-variables forme un nuage de variables qui vit dans un espace à I dimensions.

Nous avons vu quels étaient les tableaux de données auxquels s'intéresse l'ACP, quelles étaient les problématiques de l'ACP, nous verrons dans les prochaines vidéos comment mettre en œuvre l'ACP.

Deuxième partie. Etude des individus et des variables

(Diapositives 10 à 25)

Diapositive 10 (plan) :

Maintenant que nous avons défini les données sur lesquelles appliquer l'ACP et quelles étaient les problématiques de l'ACP, voyons comment mettre en œuvre l'ACP.

Diapositive 11 :

On a dit qu'un individu était une ligne du tableau et donc un point dans un espace à K dimensions. Si K=1, s'il n'y a qu'une seule variable et il est facile de représenter les individus. On peut faire une représentation axiale et positionner les individus en fonction de la valeur qu'ils prennent pour la variable.

Avec 2 variables, on peut construire un nuage de points classique, comme avant une régression simple, avec une première variable sur l'axe des X, la 2ème variable sur l'axe des Y.

Avec trois variables, la représentation des individus est plus difficile mais certains logiciels permettent de visualiser des points en 3 dimensions. En fait l'image est en 2 dimensions mais en bougeant le nuage on peut avoir une idée de sa forme en 3 dimensions.

Mais avec 4 variables ou plus, il est impossible de représenter un nuage de points en 4 dimensions. C'est impossible à représenter et même impossible à imaginer. Par contre le concept mathématiques est simple : nous avons K cordonnées sur K dimensions, K étant plus grand que 4.

Maintenant comment faire pour visualiser le nuage d'individus si celui-ci vit dans un espace très grand ? Pour répondre à cette question, nous devons d'abord définir la notion de ressemblance : quand est-ce que 2 individus se ressemblent ? Deux individus se ressemblent s'ils prennent des valeurs proches sur l'ensemble des K variables. On peut définir la distance entre 2 individus. Grâce à Pythagore, la distance au carré entre 2 individus est la somme des carrés des écarts sur chacune des variables.

C'est une définition simple de la notion de ressemblance et donc finalement quand je veux étudier le nuage des individus, cela revient à étudier la forme de ce nuage : voir les individus qui sont proches, les individus qui sont éloignés.

Diapositive 12 :

L'étude du tableau de données peut être réalisée géométriquement via l'étude des distances entre individus. Et l'analyse des distances entre individus revient à étudier la forme du nuage de points. On voit sur cette photographie un vol d'étourneaux. Le nuage d'oiseaux est en 3 dimensions et la photographie permet de visualiser le nuage dans un espace plus petit, en 2 dimensions seulement. A partir de cette représentation en 2 dimensions, on a une bonne idée de la forme du nuage dans l'espace complet et donc une bonne idée des distances entre chaque oiseau. Donc quand les individus vivent dans un espace de grande dimension, on va chercher à étudier la forme du nuage et à visualiser ce nuage en 2 dimensions.

Diapositive 13 :

Pour étudier la forme du nuage, nous allons parler de deux prétraitements possibles sur les données: le centrage et la réduction.

Centrer les données revient à translater le nuage ce qui ne modifie absolument pas sa forme. Puisqu'on étudie uniquement la forme du nuage et que cette forme reste inchangée, on centrera toujours les données.

Diapositive 13 bis :

Voyons maintenant si nous devons réduire ou non les données. Dans le schéma de droite, on a une représentation de la taille d'individus en fonction de leur poids. La taille est exprimée en centimètres et le poids en quintal. Notons que les données ont été centrées. On voit que le nuage de points est très allongé et vertical.

Sur le schéma du milieu, la taille est cette fois représentée en mètres et le poids en kilogrammes. La forme du nuage est très différente puisqu'on a un nuage de points très allongé et horizontal.

Diapositive 13 ter :

Or nous avons dit que nous voulons étudier la forme du nuage de points. On voit bien que, selon l'unité choisie, centimètres et quintal ou mètre et kilogramme, la forme du nuage n'est pas du tout la même. Comment gérer cela ? Une idée est de centrer-réduire les données car si on centre et réduit les données, les variables deviennent comparables car elles sont sans unité. La réduction, on parle aussi de standardisation ou encore de normer les données, est indispensable si les unités de mesure sont différentes d'une variable à l'autre.

Maintenant si toutes les variables sont exprimées dans la même unité de mesure, il est possible de réduire ou non. La réduction conduit à accorder la même importance à chaque variable. Et ne pas réduire donne plus d'importance aux variables qui ont une variabilité plus grande, une variance plus grande. L'importance d'une variable sera proportionnelle à son écart-type. Donc il y a une discussion à avoir pour normer ou non, c'est-à-dire réduire ou non avant l'ACP.

Dans la suite du cours, on va toujours centrer et réduire les variables du jeu de données.

Diapositive 14 :

Voici le jeu de données de l'exemple centré-réduit. Ce tableau de données centré-réduit donne déjà quelques informations : on voit par exemple qu'à Brest, en juillet on a une valeur de -2.06. Cela signifie que les températures à Brest sont assez extrêmes en juillet. La valeur de -2 signifie que ce sont des valeurs inférieures à la moyenne et assez extrêmes. On sait que si les valeurs d'une variable suivent une loi normale alors 95 % des valeurs centrées-réduites sont comprises entre -1.96 et 1.96. Ici on ne sait pas si des données suivent une loi normale mais une valeur centrée-réduite de -2 est très extrême. Donc à Brest, il fait particulièrement froid en juillet. A Nice, en revanche, il fait particulièrement chaud surtout en février et octobre-novembre. Les températures sont supérieures à la moyenne et particulièrement grandes.

L'ACP, l'analyse en composantes principales va permettre d'analyser ce tableau centrée-réduit. Alors analyser ce tableau centré réduit, ça veut dire le visualiser. C'est difficile de voir ce tableau puisqu'il est en 12 dimensions. On ne peut pas le visualiser en 12 dimensions, on va chercher à en avoir une image approchée. Mais on va chercher une image qui s'en rapproche le plus possible.

Diapositive 14 bis :

On va donc ajuster ce nuage des individus, c'est-à-dire l'approcher au mieux. L'objectif de l'ACP sera de fournir une image simplifiée du nuage de points qui soit la plus fidèle possible, qui permet d'avoir une idée des distances entre individus, et que ces distances entre individus soient les plus fidèles possible. Donc l'ACP va chercher un sous-espace qui résume au mieux les données.

Résumer au mieux les données, ça veut dire quoi ? L'image est bonne si elle restitue le plus fidèlement possible la forme du nuage original.

On peut voir ça sur une petite animation 3D. Alors on a un nuage de points en 3 dimensions. Je vais bouger ce nuage de points. Le fait de bouger le nuage de points permet d'avoir une idée de sa forme en 3D. Seulement, on cherche uniquement une représentation en 2 dimensions, uniquement un plan. Voilà une première proposition de la visualisation du nuage 3D. Là une deuxième proposition et enfin une troisième proposition du nuage 3D. Et donc la question est de savoir parmi ces 3 propositions quelle est celle qui restitue au mieux le nuage dans son ensemble ? J'ai droit à une seule photographie de mon nuage 3D, laquelle choisir ? On va avoir tendance à choisir la troisième proposition. Pourquoi ? Parce qu'on a bien séparé les points, ce qui nous donne l'impression de mieux voir les distances entre individus. C'est l'intuition qui nous dit de bien séparer les points. Et cette intuition est bonne.

Par conséquent, une image est bonne si elle restitue fidèlement la forme du nuage : autrement dit, une image est bonne si on visualise bien la diversité, la variabilité, qu'il y a dans les données et surtout une image est bonne si elle ne déforme pas trop les distances entre individus. Les distances entre individus sont au départ calculer dans un espace à 12 dimensions et en les plongeant dans un espace à 2 dimensions elles ont été un peu déformées. On cherche à ce qu'elles soient le moins déformées possible.

Diapositive 15 :

Alors comment peut-on quantifier la qualité d'une image ? Tout simplement avec la notion de dispersion. Plus un nuage de points sera dispersé mieux on verra le nuage. Et un nuage de points est très dispersé s'il y a une forte variabilité. Cette variabilité est sur plusieurs dimensions et une variabilité sur plusieurs dimensions c'est ce qu'on appelle l'inertie. On va donc parler d'inertie maintenant. L'inertie est une variance mais généralisée à plusieurs dimensions.

Diapositive 16 :

Voici une petite illustration, là encore d'une visualisation d'un nuage 3D. Ici nous avons un animal qui est en réalité en 3 dimensions mais nous avons une photographie qui le représente uniquement sur un plan en 2 dimensions. Et quelle est l'image qui restitue au mieux la forme du nuage dans l'espace global ? C'est plutôt l'image de droite. On reconnaît un chameau parce qu'on voit beaucoup plus de choses et notamment 2 bosses et 4 pattes.

Diapositive 16 bis :

Décrivons plus en détail comment ajuster le nuage des individus, c'est-à-dire comment trouver la meilleure image approchée du nuage. En fait, on va le faire en plusieurs étapes. On va commencer par trouver le 1er axe, on parle aussi de 1er facteur, qui déforme le moins possible le nuage de points. Donc si on note H_i la projection d'un individu sur un axe, O le centre de gravité du nuage, iHi^2 est l'écart entre l'individu i dans l'espace initial et sa projection sur un axe. Et on veut que l'écart entre un individu et sa projection soit le plus petit possible. Plus précisément, on veut que le carré de cet écart soit le plus petit possible. Comme la distance Oi d'un individu à l'origine est toujours la même, iH_i est petite si OH_i est grande (ça c'est grâce à Pythagore). On sait que $OH_i^2 + iH_i^2 = Oi^2$. Donc comme Oi est constant, iH_i^2 est petit si OH_i^2 est grand. Donc on veut avoir les OHi^2 les plus grands possible. Comme je veux ça sur tous les individus, je veux que la somme des OH_i^2 soit la plus grande possible. Et la somme des OH_i^2 c'est bien la dispersion des individus la plus grande possible. On trouve ainsi le premier facteur ; ensuite on va chercher le meilleur plan.

Donc trouver le meilleur plan revient à trouver les points Hi qui appartiennent, non plus à un axe, mais qui appartiennent à un plan, et tels que la dispersion soit la plus grande possible là encore. Donc il faut que la somme des OH_i^2 soit la plus grande possible, que les points soient les plus dispersés possible. Alors une remarque : le meilleur plan, donc le plan qui permet de visualiser au mieux le nuage de points, contient le meilleur axe.

Pour trouver le meilleur plan, on va chercher le meilleur axe et ensuite chercher un axe orthogonal à ce premier axe et qui maximise la somme des OH_i^2, autrement dit qui maximise l'inertie. Puis une fois qu'on a trouvé le 2ème axe on peut chercher un 3ème axe et donc séquentiellement chercher les axes les uns après les autres. Et à chaque fois un axe doit être orthogonal à tous les axes précédents et maximiser l'inertie.

Diapositive 16 ter :

Visualisons cela avec cette animation. Ici j'ai une théière en 3 dimensions et je vais trouver les axes qui permettent de visualiser au mieux cette théière. Donc je trouve le premier axe. C'est celui-ci qui permet de visualiser au mieux la théière. Si on avait à projeter les points de la théière sur un axe on aurait pris l'axe en rouge. Et maintenant on cherche un 2ème axe orthogonal au 1er. On fixe la première dimension et on cherche un axe orthogonal au 1er qui permet de visualiser au mieux la théière. Et donc là on a trouvé le 1er axe en rouge et le 2ème axe en vert.

Diapositive 16 quater :

Alors sur nos données, avec les 15 villes, les 14 variables quantitatives, donc les 12 variables de température moyenne et les deux variables géographiques, latitude et longitude. On va visualiser le nuage des individus uniquement à partir des données de températures. Donc on s'intéresse, dans un premier temps, uniquement aux 12 premières variables de température.

Diapositive 17 :

La meilleure représentation des individus est la suivante. Ce graphe montre par exemple que Montpellier et Marseille sont très proches. Que signifie Montpellier et Marseille sont très proches ? Cela signifie que les températures moyennes à Montpellier et à Marseille sont à peu près les mêmes,

et ce, quel que soit le mois de l'année. Les valeurs sont à peu près les mêmes donc les points sont proches.

De même, Rennes et Nantes sont deux villes ayant des températures proches pour les douze mois de l'année. Par contre, Nice et Lille ont des comportements très différents. Ces 2 villes sont complètement opposées, opposées sur le premier axe. Donc si elles sont opposées sur le 1er axe, cela signifie que ce sont des villes très différentes puisque le premier axe est celui qui sépare au mieux les points. Donc l'axe qui sépare au mieux les points est un axe qui sépare Lille et Nice et donc ces deux villes ont des comportements très différents, et ce, sur l'ensemble des variables.

Diapositive 18 :

Alors maintenant qu'est-ce qui sépare, qu'est-ce qui oppose Lille à Nice ? Pour répondre à cette question, on peut avoir une bonne connaissance des données et dire qu'à Lille il fait plutôt froid et à Nice plutôt chaud mais si on veut raisonner uniquement à partir du jeu de données, on va vouloir utiliser les variables pour interpréter ces dimensions de variabilité.

On peut par exemple colorier les individus en fonction de la valeur qu'ils prennent pour une variable, ici la température en octobre. L'échelle de couleur va du bleu pour les valeurs faibles au rouge pour les valeurs élevées. On voit alors que les villes à gauche du graphe sont en bleu, et il y fait donc froid en octobre (par rapport aux autres villes), tandis que les villes à droite sont en rouge, et il y fait donc chaud en octobre. On visualise ainsi une corrélation entre les coordonnées des individus sur l'axe horizontal et la variable octobre. Comment faire maintenant, si on a beaucoup de variables, pour détecter rapidement les variables les plus intéressantes pour expliquer les dimensions sans avoir à construire de nombreux graphes ?

Qu'est-ce qui oppose Lille à Nice ? Pour le savoir, on va considérer les coordonnées des individus sur les axes. Par exemple ici l'individu Brest. On va récupérer sa coordonnée sur le 1er axe, et la noter F_{i1}, 1 pour 1er axe, et puis sa coordonnée sur le deuxième axe, F_{i2}. Donc pour chaque individu nous collectons sa coordonnée sur l'axe horizontal et sa coordonnée sur l'axe vertical. On peut ainsi créer 2 vecteurs, un 1er vecteur qui regroupe toutes les coordonnées de tous les individus sur le 1er axe, et un deuxième vecteur qui regroupe toutes les coordonnées de tous les individus sur le deuxième axe. Par construction, ces vecteurs ont I coordonnées, le même nombre que chaque variable du tableau de données.

Diapositive 19 :

Pour interpréter le graphe des individus, on peut calculer la corrélation entre la variable, par exemple la variable janvier, et l'axe 1. On peut aussi calculer la corrélation entre la variable janvier et l'axe 2. Si la variable janvier est très liée à l'axe 1, cela voudra dire que les températures en janvier sont très liées aux coordonnées sur l'axe 1. Ainsi, si la corrélation est proche de 1, cela veut dire que les individus qui ont de faibles valeurs pour janvier prennent de faibles valeurs sur l'axe 1. Les individus qui prennent de faible valeurs sur l'axe 1 sont les individus à gauche du graphe. Et les individus qui ont de fortes valeurs en janvier prennent de fortes valeurs sur l'axe 1; et seront donc à droite sur le graphe. Si la corrélation est négative, alors les individus qui prennent de faibles valeurs pour janvier prendront de fortes valeurs sur l'axe 1. Et les individus qui prennent de fortes valeurs pour janvier prendront de faibles valeurs sur l'axe 1.

Et même chose avec l'axe 2. On a la corrélation avec l'axe 2. Et donc on va pouvoir construire un graphe avec une représentation de toutes les variables du jeu de données. Toutes ces variables vont se retrouver dans un cercle qu'on appelle le cercle des corrélations, et on appelle ce graphe, le graphe du cercle des corrélations.

Diapositive 20 :

Pour l'exemple, le graphe du cercle des corrélations est le suivant: on voit que toutes les variables sont corrélées à l'axe 1, au facteur 1. On a une corrélation qui est positive et supérieure à 0.6 - 0.7 pour toutes les variables. On a même des corrélations très élevées pour octobre et mars par exemple, des corrélations très proches de 1. Cela signifie que les températures en octobre sont très liées aux coordonnées sur l'axe 1. Autrement dit, les villes qui sont à gauche, avec une faible coordonnée sur l'axe 1 ont des températures faibles au mois d'octobre, les villes qui sont au milieu ont des températures moyennes en octobre et les villes qui sont à droite ont des températures élevées en octobre. Quand je dis des températures élevées, c'est par rapport aux autres villes, le même mois de l'année.

C'est ce qu'on avait vu sur notre graphe des individus habillés en fonction de la variable. Si maintenant on dessine les individus en fonction de la valeur qu'ils prennent pour la variable juin, on visualise une évolution des températures en juin avec des villes froides en haut à gauche et des villes chaudes en bas à droite. Cette évolution des températures est exactement donnée par la représentation de la variable juin sur le graphe des variables.

Donc comment peut-on interpréter l'axe 1? En fait on a toutes les variables qui sont très liées à l'axe 1. Donc, à droite du graphe, on a toutes les villes pour lesquelles il y a de forte valeurs à la fois pour janvier, décembre, février, novembre, etc. donc pour tous les mois de l'année. A droite, les villes qui ont une forte coordonnée sur l'axe 1 sont des villes où il fait plutôt chaud tous les mois de l'année. Et à gauche, on a des villes où il fait plutôt froid tous les mois de l'année. Et ça c'est le principal facteur de variabilité. Ce qui différencie le plus les villes, c'est qu'il y a des villes où il fait froid un peu tout le temps et d'autres où il fait chaud un peu tout le temps.

Maintenant qu'est-ce qui sépare les villes par rapport au deuxième axe ? Dans le 2ème axe, les corrélations sont un peu moins fortes. Ceci est normal puisque c'est un axe de variabilité qui est moins important c'est le 2ème axe de variabilité. Donc en haut du graphe des individus, on a des villes où il fait plutôt chaud en janvier décembre, et plutôt froid en mai, juin, juillet. Il fait plutôt chaud en janvier décembre parce que la corrélation avec l'axe 2 est positive. Elle n'est pas très proche de 1 mais elle est positive (de l'ordre de 0.5) tandis que la corrélation est négative avec les variables mai juin juillet. On peut donc dire que les villes qui ont des coordonnées plutôt élevées sur l'axe 2 vont prendre des valeurs plutôt faibles en mai, juin, juillet. Autrement dit, en haut du graphe, on va avoir des villes où il fait plutôt doux l'hiver, plutôt chaud l'hiver et plutôt froid l'été. Et au contraire, les villes qui sont en bas du graphe sont des villes ou il fait plutôt chaud l'été et plutôt froid l'hiver.

Diapositive 20 bis :

On a donc déterminé les principaux facteurs de variabilité : le premier axe sépare les villes chaudes des villes froides. Le deuxième axe est orthogonal au premier et donc il différencie des villes à une

température moyenne annuelle constante. Ce 2ème axe sépare, en haut du graphe, des villes où il fait plutôt chaud l'hiver et froid l'été donc ayant une petite amplitude thermique annuelle, aux villes, en bas du graphe, où il fait plutôt froid l'hiver et chaud l'été donc aux villes ayant une forte amplitude thermique annuelle. C'est ça qui va différencier principalement les villes. Donc il y a un 1er axe de variabilité : villes chaudes - villes froides, le deuxième axe plutôt par rapport à l'amplitude thermique annuelle.

Diapositive 21 (plan) :

Nous avons vu l'étude des individus, voyons maintenant l'étude des variables.

Diapositive 22 :

Rappelons qu'une variable est un point dans un espace à I dimensions puisqu'il y a I individus. C'est donc un vecteur de I coordonnées. On va s'intéresser aux nuages des variables qui vit dans cet espace à I dimensions. Les variables seront représentées par des flèches, ce qui est la coutume en analyse en composantes principales car, comme nous le verrons, nous interpréterons surtout des angles. Donc je représente la variable k par une flèche qui part de l'origine. Le cosinus de l'angle θ_{kl} entre la flèche qui représente la variable k et la flèche qui représente la variable l, le cosinus de cet angle est égal au produit scalaire entre la variable k et la variable l divisé par la norme de la variable k et la norme de la variable l. C'est donc égal à la somme sur tous les i des $X_{ik} * X_{il}$, divisée par la racine carrée de la somme des X_{ik}^2 que multiplie la somme des X_{il}^2.

Diapositive 22 bis :

Puisque les X sont centrés, on peut lire la somme des $(X_{ik} - \overline{X}_k)$ fois la somme des $(X_{il} - \overline{X}_l)$, divisé par l'écart-type de la variable k multiplié par l'écart-type de la variable l. Donc quand les données sont centrées, on reconnaît finalement le coefficient de corrélation entre la variable k et la variable l. On a ainsi une représentation géométrique ici du coefficient de corrélation entre deux variables k et l. La représentation géométrique de cette corrélation est le cosinus de l'angle entre les variables k et l.

Diapositive 22 ter :

Si les variables sont réduites, la longueur des flèches est égale à 1 et donc l'extrémité de chaque flèche sera à une distance 1 et toutes les extrémités des flèches seront sur une hypersphère de rayon 1.

Diapositive 23 :

Comment faire maintenant pour visualiser ce nuage de variables ? Ce nuage vit dans un espace à I dimensions et on ne peut pas le visualiser sur les I dimensions.

On va utiliser la même stratégie que pour les individus, c'est-à-dire qu'on va ajuster le nuage des variables afin de visualiser au mieux le nuage des variables. On va alors chercher des dimensions qui permettent de voir au mieux le nuage des variables. Comme pour les individus, on va chercher des axes orthogonaux qui permettent de représenter au mieux les variables. Le premier axe, l'axe qui permet de voir au mieux l'ensemble des variables est l'axe qui maximise la somme des corrélations

entre le facteur et chacune des variables. Donc, le meilleur facteur, V_1, est le facteur qui est le plus lié à l'ensemble des variables. Plus lié au sens des corrélations au carré. Donc la variable V_1, sera une variable synthétique qui résume au mieux l'ensemble des variables. Cette variable 1 va porter le 1er axe et une fois que cette variable V_1est déterminée, on cherche un 2ème axe, orthogonal au premier et qui permet de bien synthétiser le reste de l'information qui n'a pas encore été synthétisé par le 1er axe. On va donc séquentiellement chercher un axe orthogonal aux axes précédents et qui maximise l'information qui n'est pas encore résumée par les premiers axes.

Diapositive 24 :

Et donc l'ajustement du nuage des variables sur notre jeu de données, eh bien c'est la même représentation des variables que la représentation qu'on avait précédemment. Tout à l'heure on avait construit une représentation des variables pour nous aider à interpréter le nuage des individus. Et maintenant, quand je construis des axes pour voir au mieux le nuage des variables, je retombe sur la même présentation.

Diapositive 24 bis :

Donc c'est un peu magique ! Incroyable ! Mais c'est ce qui fait toute la force de l'analyse en composantes principales.

Diapositive 24 ter :

On a la même représentation que précédemment. Et donc cette représentation du nuage des variables, nous a permis de caractériser le nuage des individus, la représentation des individus; elle nous a servi à caractériser les individus donc dans l'exemple, les villes froides, les villes chaudes, les villes avec une faible amplitude thermique, une forte amplitude thermique. C'est aussi, on vient de le voir ici, une représentation optimale du nuage des variables. Mais c'est également une visualisation de l'ensemble des corrélations entre les variables prises 2 à 2, visualisation grâce aux cosinus des angles entre les variables. C'est donc aussi une visualisation de la matrice des corrélations. C'est assez magique parce que c'est la même représentation.

Diapositive 25 :

Alors justement, pour les représentations, on a vu que le coefficient de corrélation entre la variable B et la variable A, est égal au cosinus de l'angle entre la variable A et la variable B. Attention, cette égalité est vraie pour l'angle dans l'espace. Or on ne visualise pas l'angle dans l'espace, mais on le visualise dans le plan de projection.

Le schéma suivant montre, à gauche, les variables dans l'espace avec le plan de projection en grisé, et à droite le plan avec leurs projections. Par exemple, les variables D et E ici sont très proches du plan de projection. Ces variables sont proches de leur projection respective, le projeté H_d est très proche de D parce que la variable D est proche du plan de projection. Et donc l'angle dans l'espace global entre D et E est très proche de l'angle entre les projections H_d et H_e, angle que l'on peut visualiser dans le plan de projection. Et donc on peut utiliser l'angle entre les projetés pour visualiser

la corrélation entre D et E. Plus exactement, le cosinus de l'angle H_dH_e est très proche du cosinus de l'angle entre D et E, parce que les variables sont bien projetées, et donc ce cosinus de l'angle entre H_d et H_e est très proche du coefficient de corrélation entre D et E.

Par contre, quand les variables sont mal projetées, ce qui est le cas par exemple pour les variables A et B, l'angle ici dans le plan de projection est petit alors que l'angle dans l'espace est très grand. Il y a une flèche qui part avec une cordonnée élevée sur une 3ème dimension quand la flèche B est plus vers le bas avec une coordonnée faible sur la 3ème dimension. Et donc, ces deux flèches se projettent à peu près au même endroit dans le plan de protection mais elles sont mal projetées et donc le cosinus de l'angle ici, est très différent du cosinus de l'angle dans l'espace. Par conséquent, on ne peut pas lire le coefficient de corrélation pour les variables A et B. Le cosinus de l'angle ici, qui est proche de 1 car l'angle est petit, n'est pas proche du coefficient de corrélation. Donc je ne peux lire les coefficients de corrélation QUE pour les variables qui sont bien projetées. Et les variables sont bien projetées si elles sont proches du cercle des corrélations.

En effet, comme on a une hypersphère de rayon 1 qui est coupée par un plan alors elle se projette dans un cercle de rayon 1. Si la flèche est proche du bord du cercle, la variable est bien projetée et donc je peux visualiser l'angle entre les variables qui sont bien projetées. Par contre pour des variables mal projetées comme A et B, leur projeté Ha et Hb se projettent à peu près au même endroit, mais il est impossible, juste avec la projection, de savoir si la flèche A est vers nous et la flèche B derrière nous ou si les 2 flèches sont vers nous. Dans un cas, on a une forte corrélation positive, dans l'autre il n'y a pas de corrélation. Donc il est impossible d'interpréter. On ne peut interpréter que les variables bien projetées.

Nous avons vu comment visualiser le nuage des variables et le nuage des individus, nous verrons dans la prochaine vidéo différentes aides à l'interprétation très utiles pour compléter l'analyse des résultats d'une ACP. Vous pouvez maintenant faire le quiz pour vous assurer que vous avez bien compris comment fonctionne l'analyse en composantes principales.

Troisième partie. Aides à l'interprétation

(Diapositives 26 à 35)

Nous avons vu dans la vidéo précédente comment construire un graphe des individus et un graphe des variables, voyons différentes aides à l'interprétation très utiles pour compléter l'analyse des résultats d'une ACP.

Diapositive 26 (plan) :

Ces aides à l'interprétation sont très utiles et sont communes aux différentes méthodes d'analyse factorielle.

Diapositive 27 :

Une première aide à l'interprétation importante est la qualité de la projection. Cette qualité de la projection, on peut la mesurer avec le pourcentage d'inertie expliquée par chaque axe ou expliquée par un plan. Le pourcentage d'inertie c'est aussi le pourcentage d'information expliquée par un axe ou un plan. On peut construire un diagramme en barres des pourcentages d'inertie expliquée par chaque dimension.

On voit ici le pourcentage d'inertie expliquée par la 1ère dimension qui est de l'ordre de 80%, et par la 2ème dimension qui est de l'ordre de 19%. Les axes étant orthogonaux, on peut additionner les pourcentages d'inertie de plusieurs axes. Ainsi, la 1ère et la 2ème dimension vont expliquer 99% de l'information qui est contenue dans le jeu de données. Cela signifie que si on résume les 12 variables initiales du jeu de données par 2 dimensions, alors on récupère 99% de l'information contenue dans tout le tableau. Autrement dit, nous avons un excellent résumé qui synthétise presque parfaitement les 12 variables. Et donc bien entendu, sur les dernières dimensions il y a très peu d'information, il n'est pas nécessaire d'aller visualiser les dimensions suivantes.

Dans d'autres exemples, il y aura beaucoup plus d'informations sur les axes 3 et 4 et il sera intéressant d'aller regarder ces axes et ces dimensions, et chercher aussi à interpréter ces dimensions. Alors ces pourcentages d'inertie, on les voit sur les graphes. C'est ce qui est écrit ici 80% d'information sur le 1er axe et 19% sur le 2ème axe.

Diapositive 28 :

Maintenant comment savoir si le pourcentage d'information est important ou pas ? On peut lire dans ce tableau le quantile à 95 % du pourcentage d'inertie du premier plan de l'ACP quand toutes les variables sont indépendantes.

Par exemple, avec 15 individus et 12 variables cette valeur de 47.8 me dit que, après avoir fait 10000 ACP avec 15 individus et 12 variables indépendantes, avoir calculer le pourcentage d'inertie du premier plan de chacune de ces 10000 ACP, 95% des valeurs sont inférieures à 47.8. Donc si le pourcentage d'inertie sur le 1er plan d'une ACP à 15 individus et 12 variables est inférieur à 47.8, alors cela signifie que le plan d'ACP n'explique pas plus que ce qu'on aurait eu avec des variables indépendantes. Cela signifie que le résumé n'est pas très synthétique.

Dans notre exemple, on avait une valeur de 98 % qui est bien supérieure à ce 47.8 et donc cela signifie que le plan de l'ACP de notre exemple résume bien l'information. Il n'y a pas que des variables indépendantes dans le jeu de données mais il y a des liaisons fortes.

Diapositive 29 :

Ici on trouve le tableau quand le nombre de variables est plus important.

Diapositive 30 :

Présentons une autre aide à l'interprétation très utile : ce sont les variables supplémentaires ou illustratives. Les variables supplémentaires sont des variables qui ne servent pas à construire les axes, c'est-à-dire que ces variables ne servent pas à calculer les distances entre individus. En revanche, ces variables peuvent être utilisées en complément, en supplémentaire, pour aider à interpréter les axes. On peut avoir des variables quantitatives ou des variables qualitatives en supplémentaire.

Pour les variables quantitatives, c'est très simple, on va projeter ces variables quantitatives sur le plan de l'ACP exactement comme on avait projeté les variables pour nous aider à interpréter le graphe des individus. Par exemple pour la variable latitude, on calcule le coefficient de corrélation entre la latitude et la 1ère dimension et le coefficient de corrélation entre la latitude et la 2ème dimension et on dessine la variable.

Et donc avec cette projection, on a la latitude, la longitude puis 2 variables qu'on avait envie de regarder une fois qu'on avait commencé à interpréter nos données. On avait vu que le premier axe semblait opposer des villes où il faisait plutôt froid tout le temps et des villes où il faisait plutôt chaud tout le temps, et donc on a calculé la variable "température moyenne annuelle". On voit que cette variable "moyenne annuelle" est très corrélée à la 1ère dimension. Le coefficient de corrélation avec la 1ère dimension est égal à 1 (en arrondissant à 2 décimales). Cela signifie que la 1ère dimension c'est la moyenne annuelle.

La 2ème dimension est, quant à elle, très liée à l'amplitude thermique annuelle avec une corrélation qui est proche de -1, c'est-à-dire que les villes qui ont une faible coordonnée sur la 2ème dimension, comme Lyon, par exemple ont une forte amplitude annuelle et les villes qui sont en haut, par exemple Brest, sont des villes avec une faible amplitude thermique annuelle. Donc ça c'est pour les variables quantitatives supplémentaires.

Pour les variables qualitatives supplémentaires on va considérer les modalités des variables qualitatives. Et on va donc projeter chaque modalité au barycentre des individus qui prennent la modalité. La variable région prend 4 modalités nord-est, nord-ouest, sud-est, sud-ouest. On va calculer le barycentre des 3 villes du nord-ouest, Brest, Rennes, Nantes et positionner la modalité nord-ouest au barycentre de ces trois villes. De même, la modalité nord-est est au barycentre des villes Lille, Paris, Strasbourg qui sont les 3 villes du nord-est. De cette façon on va projeter toutes les modalités de la variable qualitative.

Notez bien que la variable qualitative est représentée, non pas sur le graphe des variables mais sur le graphe des individus. On visualise les barycentres des modalités.

Diapositive 31 :

Alors maintenant, deux aides à l'interprétation : la qualité de représentation et la contribution.

On va s'intéresser à la qualité de représentation d'une variable ou d'un individu. La qualité de représentation est mesurée par le cosinus au carré de l'angle entre une variable et son projeté sur le plan ou sur un axe; et de même pour un individu on va calculer le cosinus carré de l'angle entre le vecteur qui part du centre du nuage et va jusqu'au point i et le vecteur qui part du centre du nuage et va sur le projeté Hi. Ce cosinus carré est calculé axe par axe donc on voit par exemple que pour la variable janvier le $\cos^2$ avec la 1ère dimension vaut 0.58 et avec la 2ème dimension vaut 0.42. La qualité de projection sur le plan s'obtient en sommant ces deux valeurs : 0.58 + 0.42 = 1, ce qui signifie que le $\cos^2 = 1$ donc cela signifie que l'angle est très proche de 0, et donc la variable est extrêmement bien projetée.

De même, pour les individus, le cosinus carré pour Bordeaux vaut 0.95 pour la 1ère dimension, 0 pour la deuxième et donc la somme est égale à 0.95 pour le plan ; donc Bordeaux est une ville qui est extrêmement bien projetée sur le plan. Alors pourquoi s'intéresser à la qualité de représentation ? Parce que les proximités entre individus bien projetés, ou les liaisons entre variables bien projetées, vont pouvoir être interprétées. Par contre si 2 individus sont mal projetés, cela signifiera qu'ils sont loin du plan de projection et donc même s'ils sont proches sur le plan peut-être qu'ils sont éloignés dans l'espace par rapport à une 3ème ou une 4ème dimension. Donc il ne sera pas possible d'interpréter leur proximité.

Au niveau des contributions, qui est un autre critère, on peut calculer la contribution d'une variable ou la contribution d'un individu à la construction d'un axe. La contribution d'une variable correspond juste au carré de la corrélation entre la variable et l'axe, divisé par la somme des carrés des corrélations entre les variables et l'axe. Quand les variables ne sont pas réduites, il faut multiplier les carrés des coefficients de corrélation par la variance de chaque variable. Les contributions permettront de savoir si l'axe a été construit surtout à cause d'une ou quelques variables très particulières. De même, la contribution d'un individu est la coordonnée au carré de l'individu sur l'axe divisé par la somme des carrés des coordonnées de l'ensemble des individus. Ces contributions sont exprimées en pourcentage.

Là encore si un individu contribue très fortement à la construction d'un axe, cela signifie que, peut-être qu'à lui seul, il contribue à la formation de l'axe. Et donc que, sans cet individu, on aurait un axe différent. Dans ce cas-là, on va dire que l'individu est très particulier, on va expliquer en quoi l'individu est très particulier, et après, il peut être intéressant de refaire l'analyse sans cet individu pour voir si les axes changent, si la 1ère dimension de variabilité est toujours la même ou bien si elle est un peu différente sans cet individu très particulier.

Si on colorie les individus en fonction de leur qualité de représentation sur le plan, on voit que les individus sur l'extérieur du graphe semblent mieux projeter. C'est souvent le cas. Mais ici, tous les individus sont très bien représentés car l'échelle varie entre 0.9 et 1. Si on colorie maintenant les individus en fonction de leur contribution à la construction du plan, i.e. des deux premières dimensions, on voit que les individus qui sont les plus éloignés du centre de gravité sont ceux qui contribuent le plus à la construction de ce plan. La contribution est simplement fonction de la distance au centre de gravité du nuage.

Diapositive 32 :

Voici maintenant des aides à l'interprétation très importantes quand on a beaucoup de variables. C'est une description automatique des axes par les variables, avec les variables quantitatives et puis ensuite on verra avec les variables qualitatives. Alors comment faire ? On va calculer les corrélations entre chaque variable et la dimension et on va trier les coefficients de corrélation. Ensuite on conserve uniquement les coefficients de corrélation significativement différents de 0.

Pour la description de la première dimension, on voit que la variable la plus corrélée est la variable moyenne qui a une corrélation extrêmement proche de 1 et qui permet de décrire au mieux le 1er axe. Mais il y a d'autres variables, la variable octobre ou septembre qui sont extrêmement liées et sont corrélées positivement à la 1ère dimension. On voit aussi que la variable latitude est corrélée négativement à la 1ère dimension.

Donc il y a beaucoup de variables qui permettent de caractériser la 1ère dimension alors que pour la 2ème dimension est décrite par moins de variables. Ceci est attendu puisque la 2ème dimension est moins liée aux variables que la 1ère, par construction, la 1ère dimension est la principale dimension de variabilité. Les variables janvier et décembre sont corrélées positivement à cette 2ème dimension puis les variables juillet, longitude et amplitude sont corrélées négativement. On voit que la description des dimensions est possible à la fois avec des variables actives, les mois de l'année, et des variables supplémentaires longitude, amplitude, moyenne par exemple. Voilà pour les descriptions par les variables quantitatives.

Diapositive 33 :

Pour les variables qualitatives, on va utiliser la méthode d'analyse de variance. Plus précisément, on va considérer le modèle qui explique la variable Fs qui correspond aux coordonnées des individus sur l'axe s en fonction de la variable qualitative qui nous intéresse. Donc si on a plusieurs variables qualitatives, on va construire une analyse de variance par variable qualitative.

On va construire tout d'abord un test de Fisher pour voir si la liaison est significative entre la variable et les coordonnées des individus sur l'axe. Et on peut ensuite construire des tests de Student pour tester l'effet de chaque modalité et voir si chaque modalité est différente de la moyenne des modalités. Pour la variable région, le rapport de corrélation eta² de l'analyse de variance vaut 0.6, ce qui s'interprète comme "60% de la variabilité des coordonnées des individus sur la 2ème dimension est expliquée par la variable région". Et ça c'est significatif.

Et puis, plus dans le détail, on voit que les villes du nord-ouest de la France ont une coordonnée significativement positive sur le 2ème axe. Au contraire les villes du sud-est de la France ont une coordonnée significativement négative sur le 2ème axe.

Diapositive 34 :

Pour résumer, quelle est la pratique de l'ACP ?

La première chose à faire est de choisir quelles sont les variables qui seront actives, c-à-d qui vont servir à calculer les distances entre individus; quelles sont les variables qui vont servir à construire les

axes. Les autres variables ne serviront pas à construire les axes mais pourront servir à interpréter les axes, à interpréter les dimensions.

Ensuite il faut déterminer si on réduit ou non les variables. On a vu que, si les variables sont exprimées dans des unités différentes, il est nécessaire de réduire, c'est indispensable; par contre si les variables sont toutes de même unité, on peut réduire et accorder la même importance à chaque variable ou ne pas réduire et accorder plus d'importance aux variables qui ont une plus grande variance.

Ensuite on construit l'ACP, et on choisit le nombre de dimensions à interpréter : est-ce qu'on interprète uniquement 2 dimensions ou est-ce qu'on interprète aussi la 3ème, la 4ème dimension auquel cas on construira un graphique avec les dimensions 1-2 mais aussi un graphique avec les dimensions 3-4.

Ensuite on interprète les résultats en regardant simultanément le graphe des individus et le graphe des variables. On va faire des allers-retours entre ces 2 graphes pour décrire et comprendre les différences et ressemblances entre individus et pour comprendre les liaisons entre variables.

Les différents indicateurs que nous avons présentés permettent d'enrichir l'interprétation. On peut calculer des contributions, utiliser les cos^2 pour être sûr que les proximités que l'on visualise sur le plan entre individus sont bien des proximités aussi dans l'espace global.

Et finalement quand on dit que 2 individus se ressemblent, il est intéressant de revenir aux données brutes pour être sûr de ne pas faire une mauvaise interprétation. A chaque fois que l'analyse nous suggère une interprétation il faut revenir aux données brutes pour valider ce que l'on est en train d'interpréter. Il est possible de revenir aussi aux données centrée-réduites plutôt qu'aux données brutes.

Diapositive 35 :

Alors quelques compléments.

Il y a un livre sur l'analyse de données, il y a beaucoup de livres sur l'analyse des données, mais il y en a un qui reprend cette présentation du cours: L'analyse de données avec R, aux Presses Universitaires de Rennes.

Le package FactoMineR permet de faire des ACP avec des variables supplémentaires quantitatives des variables supplémentaires qualitatives avec toutes les aides à l'interprétation qui ont été décrites dans ce cours. Un site internet est dédié au package FactoMineR et des vidéos montrent comment utiliser le package pour mettre en œuvre les ACP sous le logiciel.