Chapitre 6
Analyses factorielles

Sandrine CABUT
Aix-Marseille Univ., CNRS, Minist. Culture, LAMPEA, Aix-en-Provence

Les méthodes descriptives présentées au chapitre 5 permettent la visualisation de liens entre deux variables. Elles peuvent être reproduites quel que soit le nombre de variables à étudier. Toutefois, cela nécessite de reproduire autant de tableaux, autant de nuages de points que nécessaire pour étudier chaque lien deux à deux. D'autres méthodes sont donc nécessaires pour les généraliser et étudier conjointement un plus grand nombre de variables et de liens.

Les analyses factorielles sont des méthodes de visualisation des données qui permettent d'optimiser des critères algébriques et géométriques bien définis, afin d'obtenir la meilleure représentation correspondante. Ces analyses de données ne font pas appel à des hypothèses d'ordre statistique à proprement parler. Largement utilisées dans de nombreuses disciplines d'application des statistiques, leur essor est lié à celui des moyens de calcul modernes. Les liens entre un grand nombre de variables quantitatives*, ou qualitatives*, sont ainsi visualisés aussi aisément que grâce à un nuage de points.

1 Analyse en Composantes Principales (ACP)

Pearson (Pearson, 1901) est le premier à avoir développé cette méthode, ensuite intégrée à la statistique mathématique et à l'analyse multivariée par Hotelling (Hotelling, 1933) et Anderson (Anderson, 1958).

En effet, dès lors que l'on s'intéresse à p ($p > 2$) variables quantitatives ($x^1 \cdots x^j \cdots x^p$) observées sur n individus M_1, ..., M_n, soit un tableau X de terme général x_i^j, une représentation aussi simple que le nuage de points vu précédemment n'est plus envisageable. En effet, il nous est impossible de visualiser un espace de dimension p. Une vue 3D est possible sur les logiciels actuels, mais au-delà ? L'ACP entre alors en scène. Tel un photographe aguerri, elle va chercher le meilleur angle de vue pour rendre compte de

la position relative des n individus statistiques dans cet espace à p dimensions (noté $\mathbb{R}^p$)[1], pour en prendre un cliché en 2D, et donc aisément visualisable.

$$\mathbf{X_{n\times p}} = \begin{bmatrix} x_1^1 & \cdots & x_1^j & \cdots & x_1^p \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_i^1 & \cdots & x_i^j & \cdots & x_i^p \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_n^1 & \cdots & x_n^j & \cdots & x_n^p \end{bmatrix} = [x^1 \cdots x^j \cdots x^p]$$

1.1 Principes

L'ACP peut ainsi être abordée d'un point de vue strictement géométrique. La prise de vue correspondant à la projection sur un plan, choisi de sorte à déformer le moins possible les distances, et donc à obtenir un nuage de points d'inertie*, et donc de dispersion*, maximale. Dans les faits, l'ACP effectue des « clichés » unidimensionnels sur des axes factoriels (Figure 1), que l'on peut coupler en fonction de ce que l'on souhaite observer et des résultats obtenus sur chaque axe. Un couple d'axes factoriels (V^α, V^β) (α et $\beta \leq p$, $\alpha \neq \beta$) forme un espace vectoriel de dimension 2, nommé plan factoriel, qu'on notera de façon simplifiée plan factoriel (α, β).

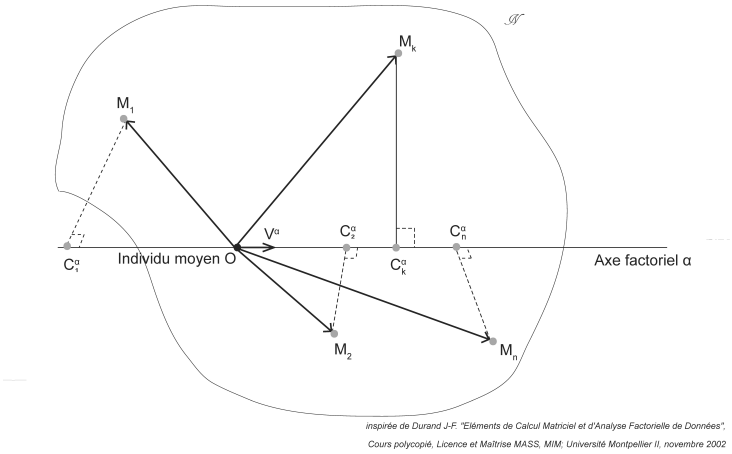


FIG. 1 : Projection sur l'axe factoriel V^α du nuage $\mathcal{N}$ des points-lignes (individus) de l'espace $\mathbb{R}^p$

De façon duale, l'ACP rend compte de la position relative des p variables dans l'espace à n dimensions (noté $\mathbb{R}^n$). Ainsi, l'ACP permet de visualiser graphiquement les liens linéaires[2] entre les p variables quantitatives en jeu.

[1] L'ensemble des $p-$uplets $(x_i^1, x_i^2, \cdots, x_i^p)$ dont chaque composante est un réel

[2] Les liens quadratiques, exponentiels ou autre ne sont pas étudiés par l'ACP.

On obtient ainsi deux représentations complémentaires, mais non superposables (car réalisées dans des espaces vectoriels différents $\mathbb{R}^p$ et $\mathbb{R}^n$, respectivement pour les individus et les variables), s'interprétant conjointement (Figure 2).

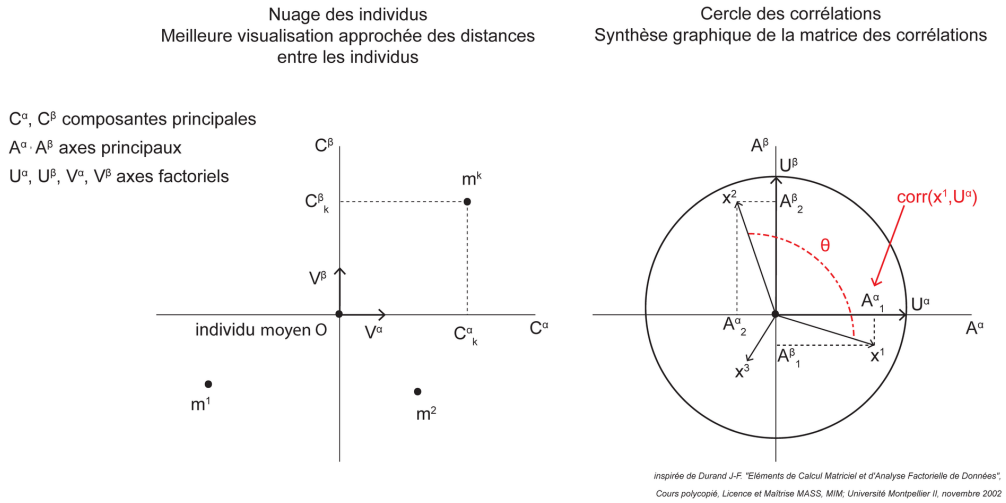


FIG. 2 : ACP - Visualisation du nuage des individus et du cercle des corrélations* sur le plan factoriel (α, β)

D'un point de vue matriciel, les axes factoriels de l'ACP sont obtenus par diagonalisation de la matrice des variances-covariances* (ACP uniquement centrée) des variables, ou de la matrice de corrélation* pour une ACP normée (i.e. centrée-réduite). On obtient ainsi les vecteurs propres v_α et valeurs propres λ_α associées, ces dernières étant ordonnées de façon décroissante et correspondant à la variance expliquée par l'axe de rang α. De nombreux ouvrages détaillent ces calculs, notamment ceux de Saporta (Saporta, 2011), Lebart (Lebart et al., 2006), et Tenenhaus (Tenenhaus, 2007), les lecteurs intéressés peuvent s'y reporter pour plus de détails.

Concrètement, k ($\leq p$) autres variables, appelées composantes principales, combinaisons linéaires des variables initiales (1.1), sont construites telles que :

- Les composantes principales, C^α étant la composante de rang α, sont ordonnées selon l'information (variance) qu'elles restituent, la 1ère étant celle qui restitue le plus d'information
- Les composantes principales sont des vecteurs orthogonaux, c'est-à-dire des variables non corrélées entre elles

$$C^\alpha = \sum_{i=1}^{p} x^i F_i^\alpha \tag{1.1}$$

où F_i^α est un **scalaire**[3] mesurant l'influence de la variable x^i sur la nouvelle variable C^α

[3] *Note des éditeurs* : un scalaire est une grandeur dont la valeur ne dépend que du point auquel on l'évalue et qui est indépendante du système de coordonnées.

En d'autres termes, on procède à un changement de repère de telle sorte que le premier axe représente la plus grande part de l'inertie du nuage, le deuxième axe représente la plus grande part de l'inertie non prise en compte par le premier axe, etc…

Pourquoi une ACP normée ?

Le **centrage**[4] est obligatoire afin que le nuage de points des individus soit centré sur l'individu moyen.
La **réduction**[5] permet de donner la même importance aux variables. Elle est obligatoire lorsque les variables n'ont pas la même unité de mesure, le même ordre de grandeur[6] ou la même dispersion. Ce qui est fréquemmment le cas. Qui plus est, cela permet de travailler sur les corrélations des variables (au lieu des covariances) et de faciliter les interprétations.

C'est dans le cadre d'une ACP normée que nous nous placerons dorénavant, le tableau de données X étant alors remplacé par la matrice X' de terme général :

$$x_i'^j = \frac{x_i^j - \overline{x^j}}{s^j \sqrt{n}} \quad \text{avec } s^j \text{ l'écart-type de la variable } x^j$$

1.2 Interprétations statistiques

Cercle des corrélations

En ACP normée, le cercle des corrélations est de rayon 1 et s'interprète, comme son nom l'indique, en termes de corrélations*. En effet, la coordonnée d'un point-variable sur un axe α correspond au coefficient de corrélation de cette variable avec A^α (variable représentée par cet axe) (Figure 2).

Le cosinus de l'angle θ (Figure 2) entre deux variables correspond au coefficient de corrélation entre ces deux variables. Ainsi les variables ayant :

- la même direction ($\theta < 90°$) sont corrélées positivement
- des directions opposées ($\theta > 90°$) sont corrélées négativement
- des directions orthogonales ($\theta = 90°$) ne sont pas corrélées

Ainsi le cercle des corrélations permet de visualiser aisément l'ensemble des corrélations des variables en jeu dans l'analyse, autrement dit la matrice de corrélation.

Remarque :

Il faut bien garder en tête que les corrélations présentées en ACP sont relatives à celles détectées dans le tableau X. Ainsi les plus fortes corrélations détectées apparaitront comme élevées sur le cercle des corrélations, même si ces corrélations sont en réalité très faibles.

[4] La moyenne de la variable est soustraite aux valeurs de chaque individu.
[5] Les valeurs de chaque individu sont divisées par l'écart-type des valeurs de la variable.
[6] *Note des éditeurs* : par exemple, pour les études de provenance, lorsque les compositions chimiques sont utilisées et que les mesures correspondent à des échelles différentes (en pourcent ou en parties par million), *cf.* 2.1.2 à 2.1.6.

De plus les corrélations peuvent être arbitrairement élevées alors que la relation n'est pas franchement linéaire (ex : courbe sinueuse autour d'une droite, deux nuages de points éloignés mais sur une même droite, etc…).

C'est pourquoi il est toujours nécessaire d'étudier la matrice des corrélations et le diagramme de dispersion avant toute conclusion (*cf.* 5). L'utilisation de méthodes multivariées ne dispense pas de décrire ses données grâce aux statistiques descriptives de base ! (*cf.* 4 et 5)

Nuage des individus

Les proximités entre individus s'interprètent en termes de similitudes vis-à-vis des variables en jeu.

De façon duale à la diagonalisation de la matrice de corrélation $X'^T X'$, pour construire les axes factoriels V^α sur lesquels sont projetés les points individus de l'espace $\mathbb{R}^p$, la diagonalisation de la matrice $X'X'^T$ permet d'obtenir les axes factoriels U^α de même valeur propre λ^α pour la projection des points variables de l'espace $\mathbb{R}^n$.
Les relations de transition existant entre les 2 espaces : $U^\alpha = \frac{1}{\sqrt{\lambda^\alpha}} X'V^\alpha$ sont à la base de l'interprétation conjointe des deux représentations obtenues, nuage des individus et cercle des corrélations. Ainsi, un individu est situé du côté des variables pour lesquelles il prend de fortes valeurs et à l'opposé de celles pour lesquelles il présente de faibles valeurs (*cf.* 2.1.7).

1.3 Choix du nombre d'axes factoriels

Un plan factoriel ne représente qu'une partie de la réalité (Figure 3) :

- Le premier plan factoriel (1,2), par construction, montre les relations les plus importantes
- Les plans factoriels suivants révèlent des associations ou oppositions plus discrètes

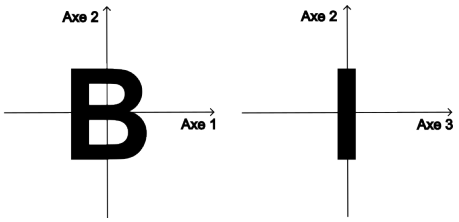


FIG. 3 : Illustration de l'information contenue par chaque plan factoriel

La valeur $I_\alpha = \frac{\lambda^\alpha}{\sum_{i=1}^k \lambda^i} \times 100$ représente le pourcentage de l'inertie totale $\sum_{i=1}^k \lambda^i$ (k étant le nombre total d'axes factoriels) restitué par l'axe α. On dit également que l'axe α exprime I_α % de la variabilité des données.
Ainsi le cumul de ces valeurs jusqu'à l'axe de rang α, soit $\frac{\sum_{i=1}^\alpha \lambda^i}{\sum_{i=1}^k \lambda^i} \times 100$, identifie le pourcentage d'information du nuage $\mathcal{N}$ exprimé par les α premiers axes principaux.

Toutefois aucun seuil universel pour ce pourcentage ne peut être admis pour choisir le nombre d'axes factoriels à retenir pour les interprétations. En effet, plus les variables sont nombreuses ou moins elles sont corrélées, plus est faible la part d'inertie restituée par les α premiers axes.

Plusieurs critères empiriques sont alors utilisés, dont les plus usités[7] :

- **Critère du coude (ou de Cattell)**

  On trace le diagramme des valeurs propres (Figure 4), *i.e.* le diagramme en barres de décroissance des valeurs propres (ou de façon équivalente des pourcentages d'inertie correspondants) en fonction de l'axe pour y déceler un décrochage. S'il existe, on conservera les axes en amont de ce décrochage. Ainsi, ici nous conservons les 3 premiers axes, qui restituent environ 77 % de l'information.

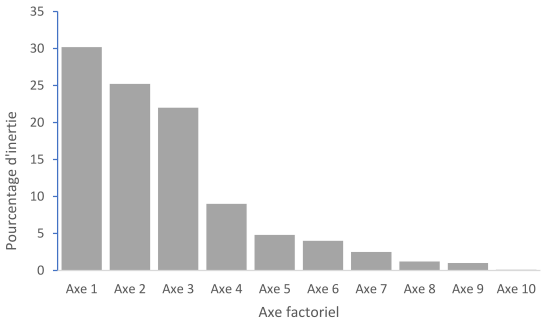


Fig. 4 : Diagramme des valeurs propres

- **Critère de Kaiser**

  On conserve les axes pour lesquels la valeur propre est supérieure à la moyenne des valeurs propres. Dans le cas d'une ACP normée, cela revient à conserver les axes dont la valeur propre est supérieure à 1.

  Méfiance toutefois dans le cas d'un facteur taille (*cf.* ci-dessous partie 1.5 pour une définition du facteur taille) très important, ce critère pouvant conduire à sous-estimer les autres dimensions.

1.4 Aides à l'interprétation

Malgré le fait que les axes factoriels aient été calculés de façon à être « les meilleurs » pour la visualisation du nuage de points et des variables, ces objets ne sont pas tous forcément bien représentés sur un plan factoriel (α, β) donné. De plus l'interprétation correcte des axes factoriels nécessite de connaitre la part que les variables ont pris dans leur constitution (via leur combinaison linéaire, formule 1.1). Des aides à l'interprétation sont donc nécessaires.

[7] De nombreux autres critères existent : intervalles de confiance d'Anderson, analyse parallèle de Horn, bootstrap, Règle de Karlis-Saporta-Spinaki, etc…

Contributions relatives d'un axe à la représentation d'un individu (ou d'une variable)

La position relative de 2 points-individus (respectivement l'angle entre 2 variables) sur un plan factoriel ne peut être interprétée que si elle reflète bien la position des points (respectivement des variables) dans l'espace $\mathbb{R}^p$ (respectivement $\mathbb{R}^n$). Prenons l'exemple des individus M_2 et M_k de la figure 1 (*cf.* ci-dessus partie 1.1). Dans l'espace $\mathbb{R}^p$ ces individus sont éloignés, pourtant leurs projections respectives sur l'axe factoriel V^α (points-individus C_2^α et C_k^α) sont proches. Cette proximité ne reflète pas la position relative réelle des individus et ne peut donc être interprétée.

La bonne représentation des individus (respectivement variables) sur un axe factoriel est donnée par le cosinus carré de l'angle formé par un point-individu (respectivement variable) avec sa projection, d'où l'appellation de $\cos^2$. La contribution relative d'un plan factoriel (α, β) à la représentation d'un individu (respectivement d'une variable) est égale à la somme des contributions relatives (*i.e.* des $\cos^2$) aux axes α et β pour cet individu (respectivement variable).

Un $\cos^2$ proche de 1 reflète une bonne représentation, un $\cos^2$ proche de 0 une mauvaise. Pour un même individu (respectivement variable), la somme des $\cos^2$ sur l'ensemble des axes factoriels est égale à 1.

Dans le cas d'une ACP normée, cette contribution est directement liée au coefficient de corrélation d'une variable avec un axe factoriel, aussi son examen n'est guère utile. En effet, la visualisation du cercle des corrélations suffit : une variable est bien représentée si elle est proche du cercle.

Contributions absolues d'un individu (ou d'une variable) à un axe factoriel

Les contributions absolues (CTA) correspondent à la part prise par chaque variable (respectivement individu) dans l'inertie expliquée par l'axe factoriel, et permettent donc d'interpréter la signification de ce dernier. Pour chaque axe factoriel, la somme des CTA, exprimées en pourcentage, de l'ensemble des variables (respectivement individus) est égale à 100 %.

Elles prennent en compte non seulement l'éloignement de l'origine mais aussi le poids d'une variable (respectivement d'un individu). Aussi, dans le cas d'une ACP normée :

- une variable de grande coordonnée sur un axe contribue fortement à cet axe
- les individus à forte CTA sont ceux les plus éloignés de l'individu moyen (*i.e.* le centre du graphique)

Les contributions des individus aux axes sont surtout intéressantes si les individus ne sont pas "anonymes", *i.e* qu'ils ont un intérêt en eux-mêmes. En effet, s'agissant par exemple de l'étude de centaines de tessons de céramique, on considère que ceux-ci sont anomymes dans le sens où peu nous importe l'interprétation du tesson numéro 112 en particulier. En revanche, dans le cas de l'étude de styles céramiques, vus comme individus statistiques (un style céramique correspond à une ligne de notre tableau X) et sur lesquels on observe différentes variables, il sera intéressant d'interpréter la position du style campaniforme en particulier.

Lorsqu'un individu présente une position extrême et donc une CTA excessive, le retirer peut fortement modifier l'analyse. Dans ce cas, et si l'individu est "anonyme", il sera préférable de l'inclure en tant qu'élément supplémentaire (*cf.* ci-dessous partie 1.6).

Il n'existe pas de valeur seuil théorique pour dire qu'une contribution ou un $\cos^2$ est faible ou fort. L'appréciation se fait empiriquement en fonction des données et des valeurs calculées (en les confrontant de façon relative les unes aux autres) et varient donc d'un jeu de données à l'autre.

1.5 Facteur taille

Lorsque toutes les variables sont positivement corrélées entre elles, le premier axe factoriel constitue un facteur taille. Toutes les variables sont alors situées du même côté de l'axe ; et cet axe restitue une très grande part de l'information. C'est généralement le cas sur les données biométriques, comme par exemple les métriques de métacarpiens de bovidés (Figure 5).

Le 2e axe factoriel différencie alors des individus de taille semblable, on l'appelle facteur de forme.

1.6 Affichage de variables ou d'individus supplémentaires (illustratifs)

Les individus et variables du tableau X sur lesquels l'ACP a été réalisée sont dits "actifs".

A contrario, les individus et/ou variables ne participant pas à l'analyse, et donc au calcul des axes factoriels, sont appelés illustratifs ou supplémentaires. Ils sont projetés dans le plan factoriel de l'ACP *a posteriori* afin de vérifier des hypothèses sur un échantillon test ou de visualiser leur lien avec les autres variables (Figure 5).

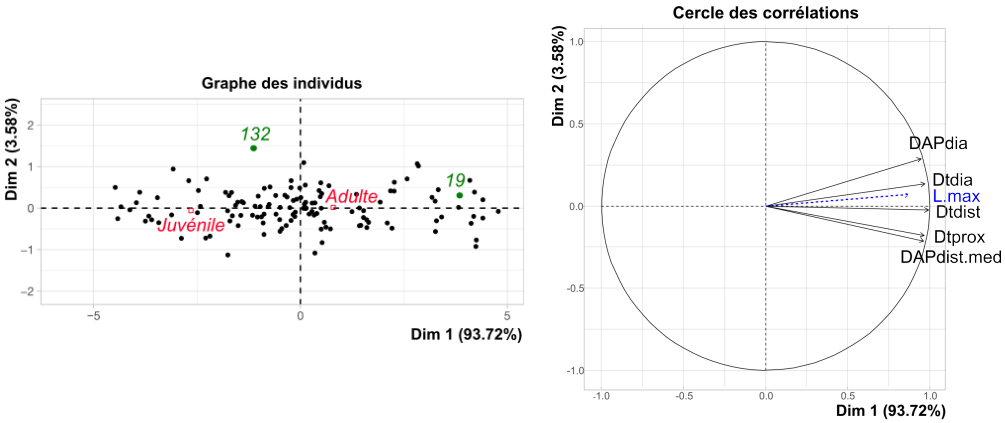


FIG. 5 : ACP sur métriques de métacarpiens de bovidés : facteur taille
Individus supplémentaires en vert ; variable qualitative supplémentaire = classe d'âge (juvénile / adulte) en rouge ; variable quantitative supplémentaire en bleu

Les individus supplémentaires x_{sup} sont positionnés aisément par simple calcul de leurs coordonnées sur les axes factoriels via la combinaison linéaire établie par l'ACP, formule (1.1) (*cf.* partie 1.1), en y injectant leurs valeurs x^j_{sup} prises sur les variables.

Ces individus illustratifs n'ayant pas participé à l'élaboration des axes factoriels, le centrage se fait par rapport aux moyennes calculées sur les individus actifs uniquement, et la réduction par rapport aux écarts-types de ces mêmes individus.

Pour les variables supplémentaires, on distingue deux cas de figure :

- **Variable quantitative**

  Le calcul des corrélations de cette variable avec les composantes principales fournit ses coordonnées sur les axes correspondants. La variable peut ainsi être représentée sur le cercle des corrélations.

- **Variable qualitative**

  Elle est constituée de m modalités vues comme m groupes d'individus qui seront traités comme individus supplémentaires. Ce sont les centres de gravité de ces groupes qui sont positionnés sur le graphe des individus.

  Pour chaque modalité k d'effectif n_k , dont le centre de gravité a pour coordonnée $G_{k\alpha}$ sur l'axe factoriel α, une valeur-test est calculée :

$$V_{\text{test}} = \frac{G_{k\alpha}}{\sqrt{\frac{\lambda^\alpha}{n_k}\frac{n-n_k}{n-1}}}$$

  Une modalité occupe une position significativement différente de la moyenne générale si en valeur absolue sa valeur-test dépasse 2 [8] : $|V_{\text{test}}| > 2$

  Sinon, cela indique que l'axe factoriel sépare mal les individus « porteurs » de cette modalité.

1.7 Intérêts de l'ACP

Cette méthode purement descriptive peut ainsi être employée pour différents objectifs :

- **Détecter les liaisons entre les variables**

  Il faut toutefois être attentif au fait que, d'une part seules les liaisons linéaires sont mises en évidence par l'ACP (basée sur les corrélations de Pearson), et d'autre part que les liaisons représentées le sont de façon relative. Ainsi un angle très aigu dans le cercle des corrélations, signifie que les deux variables en jeu sont fortement corrélées relativement aux autres corrélations existant parmi l'ensemble des variables étudiées ; mais il ne donne pas d'ordre absolu de grandeur de cette corrélation (qui peut tout à fait ne pas excéder des valeurs faibles, 0,4 par exemple). Il est donc indispensable de revenir à la matrice des corrélations avant toute interprétation archéologique.

- **Représenter assez fidèlement des individus sur un plan**, en utilisant leurs p caractéristiques.

[8] Référence à la loi de Laplace-Gauss, si n_k est assez grand.

- **Détecter des individus exceptionnels et/ou d'éventuels groupes d'individus**

  Attention toutefois, à ne pas procéder à des regroupements sur la seule base de l'ACP, cette méthode purement descriptive n'y est pas adaptée. Seules les méthodes de classification décrites au chapitre 9 peuvent permettre de confirmer ces observations. De même, aucune attribution à un groupe déjà défini (par exemple, mâle / femelle) ne sera faite sur la seule base de la position d'un individu sur un plan factoriel. Là encore il existe des méthodes statistiques adaptées à cet objectif, telles que l'analyse discriminante ou la régression logistique (*cf.* 10) par exemple.

- **Réduire les dimensions d'un problème**

  L'ACP permet de diminuer le nombre de variables étudiées (réagencées en combinaisons linéaires) sans perdre beaucoup d'information. C'est ainsi une méthode très utile avant une classification car elle permet de filtrer les fluctuations aléatoires des données en remplaçant les variables d'origines par les premiers axes factoriels (*cf.* 3.6).

- **Expliciter les groupes/classes obtenues suite à une classification**

  Nous renvoyons le lecteur à la partie archéologique de l'ouvrage pour des exemples (*cf.* 3.6, 2.2.8, 2.1.3, 2.1.4, 2.1.5).

- **Afficher des variables ou des individus supplémentaires/illustratifs**

2 Analyse factorielle des correspondances (AFC)

Présentée sous le nom d'analyse des correspondances par Benzécri (Benzecri, 1969a, 1969b), cette méthode descriptive utilise, tout comme l'ACP, des propriétés algébriques et géométriques décrites par Benzécri (Benzécri, 1973a, 1973b) et Escofier-Cordier (Escofier-Cordier, 1969). L'analyse factorielle des correspondances a pour objectif d'étudier les éventuels liens entre deux variables qualitatives nominales*. Elle s'applique à leur tableau de contingence* et cherche la meilleure représentation simultanée des profils lignes et colonnes de celui-ci.

2.1 Principes

En reprenant les notations du tableau de contingence présenté au chapitre 5 paragraphe 1 (Tableau 1), l'ensemble des profils-lignes $\mathcal{X}_i$ (respectivement colonnes $\mathcal{Y}^j$) forme un nuage de r points (respectivement s) dans un sous espace $\mathbb{R}^{s-1}$ (respectivement $\mathbb{R}^{r-1}$). Les profils moyens (correspondant aux marges lignes et colonnes) sont situés aux centres de gravité de ces nuages.
L'écart à l'indépendance est ainsi aisément visualisable. En effet, en cas d'indépendance entre les 2 variables, les profils-lignes (respectivement colonnes) sont identiques entre eux et donc identiques au profil moyen correspondant. Dans l'exemple des vases italiens du chapitre 5 paragraphe 1 (Tableau 2), cela signifierait que la région de provenance en Italie ne modifierait pas la répartition des styles de vases (et inversement). C'est donc la proximité des profils qui sera étudiée sur les plans factoriels.

Remarque :

L'AFC peut être vue comme l'ACP du tableau de contingence de 2 variables qualitatives, avec pour métrique la distance du χ^2 (*cf.* 5 formule (1.2)) :

$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} (\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}})$ distance entre les profils lignes $\mathcal{X}_i$ et $\mathcal{X}_{i'}$

$d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} (\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}})$ distance entre les profils colonnes $\mathcal{Y}^j$ et $\mathcal{Y}^{j'}$

Les individus statistiques ne sont donc pas directement représentés.

Les lignes et les colonnes du tableau de contingence jouant des rôles symétriques, les relations de transition vues en ACP permettent d'établir les relations quasi-barycentriques. Ainsi, à un coefficient de dilatation[9] près :

- les profils lignes sont barycentres* des profils colonnes
- les profils colonnes sont barycentres des profils lignes

Il est ainsi possible d'obtenir une représentation simultanée des profils lignes et colonnes sur un plan factoriel, bien qu'ils soient issus de 2 sous espaces différents (Figure 6).

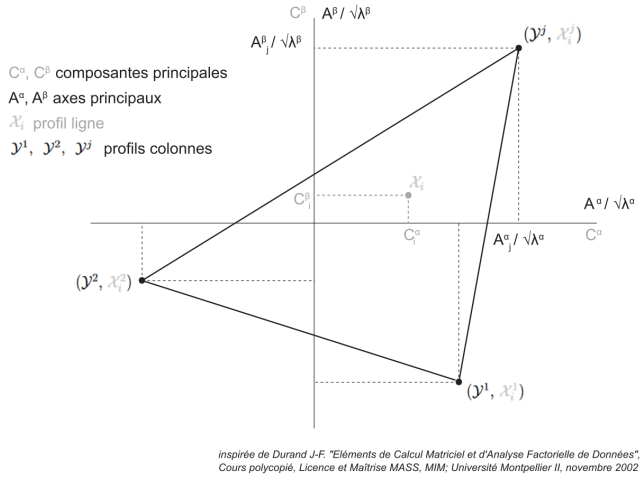


FIG. 6 : Représentation barycentrique dans le plan factoriel (α, β) du profil ligne $\mathcal{X}_i$, au barycentre des profils colonnes $\mathcal{Y}^1$, ..., $\mathcal{Y}^s$ pondérés par les valeurs du profil $\mathcal{X}_i$
Remarque : un seul profil ligne $\mathcal{X}_i$ est représenté pour ne pas alourdir la figure

[9]Le facteur de dilatation associé à un axe factoriel α est négligeable si la valeur propre λ^α est voisine de 1 ; il est de plus en plus important au fur et à mesure que l'axe factoriel est associé à des valeurs propres de plus en plus proches de 0.
Dans la pratique, on admet comme légitime la représentation simultanée des profils lignes et colonnes sur un même plan factoriel sans facteur de dilatation, bien que pour les premiers axes les valeurs propres sont souvent plus proches de 0 que de 1.

2.2 Interprétations statistiques

S'agissant de profils-lignes et colonnes, les proximités entre points représentatifs doivent être interprétées comme des proximités entre moyennes. Ainsi si deux modalités d'une même variable sont proches et bien représentées, cela signifie que leurs profils sont semblables : les deux groupes d'individus qui les possèdent ont des distributions similaires suivant les modalités de l'autre variable.

La proximité entre une modalité d'une variable et une modalité de l'autre est plus délicate à interpréter car, comme vu précédemment, elles n'appartiennent pas au même sous espace (points lignes/points colonnes). En revanche, la position d'une modalité par rapport à l'ensemble de celles de l'autre peut être interprétée : elle s'associe globalement plus à celles qui lui sont proches et moins à celles qui lui sont éloignées, et ce vis-à-vis du modèle d'indépendance.

La dispersion du nuage de points, liée à l'inertie totale, est un indicateur de la liaison entre les deux variables en jeu (Figure 7). En effet, contrairement à l'ACP et comme nous le verrons plus tard en ACM, l'inertie totale ne dépend pas uniquement du nombre de variables. Elle est liée au test classique du χ^2 et mesure l'écart à l'indépendance. Ainsi une inertie totale élevée, et donc un nuage fortement dispersé, rendra compte de l'indépendance des deux variables analysées.

Mais, il faut également s'intéresser à la forme du nuage (Figure 7). En effet, chaque axe factoriel explique une partie de l'écart à l'indépendance, quantifiée par le pourcentage d'inertie. Une direction privilégiée dans la forme du nuage sera un indicateur de liaison entre les variables (que l'inertie totale soit faible ou élevée).

Géométriquement parlant, un nuage de points profils concentrés autour du profil moyen (centre de gravité du nuage et origine du plan factoriel) selon une forme sphérique marquera l'indépendance des variables.

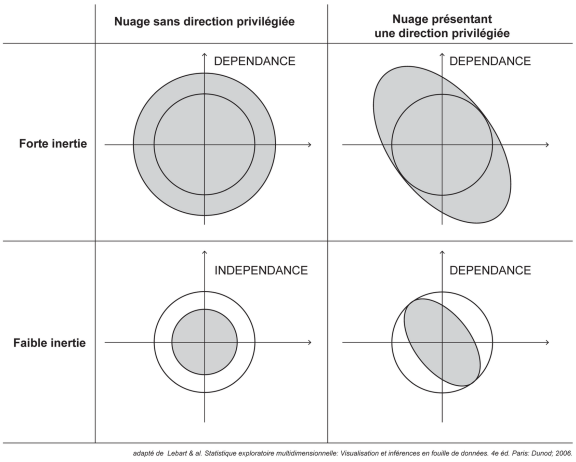


FIG. 7 : Forme du nuage de points de l'AFC et interprétation en termes d'indépendance des variables

2.3 Choix du nombre d'axes factoriels et aides à l'interprétation

Les éléments vus en ACP (*cf.* ci-dessus partie 1.3) restent valables :

- Le critère du coude est toujours valide pour définir le nombre d'axes factoriels à analyser. Leur nombre maximum étant $min\{(r-1);(s-1)\}$ où r = nombre de modalités de la variable X et s = nombre de modalités de la variable Y.
- On considère les modalités ayant les plus fortes contributions absolues comme constitutives des axes, avec pour critère :

$$CTA > \frac{n_{i.}}{n}$$

- La bonne représentation des modalités est donnée par les $\cos^2$:
  - $\cos^2$ proche de 1 : bonne représentation
  - $\cos^2$ proche de 0 : mauvaise représentation

2.4 Formes caractéristiques du nuage de points

Certaines formes du nuage de points suggèrent une possible réorganisation, par permutation des lignes et des colonnes, du tableau de données.

Décomposition en sous-nuages de points
Si toutes les valeurs propres sont proches de 1, le nuage se décompose en k sous nuages de points ($k > 2$). Chaque modalité d'une variable correspond quasi exclusivement à une seule modalité de l'autre variable, et le jeu de permutations susmentionné fait apparaitre un tableau de données en k blocs sur la diagonale (Figure 8).

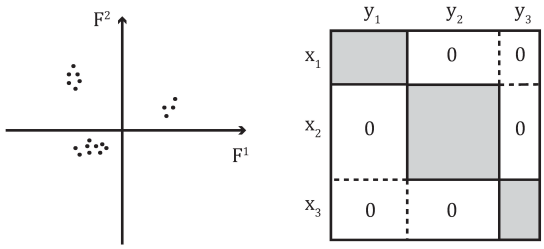


FIG. 8 : Décomposition en 3 sous-nuages de points et permutation correspondante du tableau de contingence

Effet Guttman
Le nuage de points prend une forme parabolique (Benzécri, 1973b) et le tableau de données correspondant se réorganise selon une diagonale (Figure 9). Cette réorganisation prend alors le nom de sériation (*cf.* 1.6).

Cette situation peut apparaître lorsqu'une des 2 variables est ordinale :

- Un axe, souvent le premier, oppose alors les valeurs faibles de cette variable ordinale à ses valeurs fortes

- L'autre axe du plan factoriel en jeu oppose les valeurs intermédiaires aux extrêmes

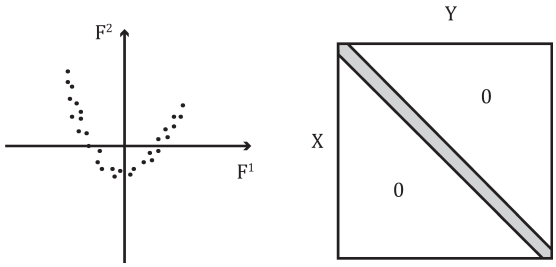


Fig. 9 : Illustration de l'effet Guttman : forme parabolique du nuage de points et tableau de contingence organisé selon une diagonale

En archéologie, la visualisation d'un tel effet, alors qu'aucune des variables en jeu ne présente d'ordre apparent, doit faire réfléchir à une interprétation en termes de structure organisée de type hiérarchique (*cf.* 2.2.8) ou chronologique par exemple (*cf.* 1.6). Une très belle illustration sur des données céramiques du Néolithique moyen figure également dans la publication de S. van Willigen (van Willigen et al., 2014).

3 Analyse des correspondances multiples (ACM)

Les principes de cette méthode remontent à Guttman (Guttman, 1941), Burt (Burt, 1950) et Hayashi (Hayashi, 1956). L'ACM est une extension de l'AFC à plus de deux variables qualitatives nominales, elle fût notamment proposée par Benzécri (Benzécri 1973a, 1973b), Escofier-Cordier (Escofier-Cordier, 1969) et Masson (Masson, 1974). A partir de p variables qualitatives $x^1, \cdots, x^j, \cdots, x^p$ (ayant respectivement $m_1, \cdots, m_j, \cdots, m_p$ modalités) et n individus, on recherche les meilleures représentations de deux ensembles constituant les lignes et les colonnes d'un tableau de données.

L'ACM peut être vue comme l'AFC du tableau disjonctif complet de ces données (*cf.* 4)[10]. Ainsi, par rapport à l'AFC, en sus des liens entre les variables, les individus sont visualisés et leurs positions vis-à-vis des modalités des p variables sont interprétables.

3.1 Principes

L'ACM permet d'obtenir deux nuages de points distincts, celui des n individus (points-lignes) et celui des $m = \sum_{i=1}^{p} m_i$ modalités (points-colonnes).

[10] En réalisant l'AFC du tableau de Burt (super tableau de contingence croisant deux à deux les p variables qualitatives), on retrouve à une constante multiplicative près les mêmes coordonnées factorielles (les valeurs propres étant toutefois différentes) (Lebart et al., 2006). Dans ce cas, comme en AFC, les individus n'apparaissent pas sur les plans factoriels calculés, et seules les modalités des variables qualitatives en jeu y figurent.

Par dérivé de l'AFC, à un coefficient $1/\sqrt{\lambda}$ près :

- la coordonnée d'un individu i est égale à la moyenne arithmétique* simple des coordonnées des modalités auxquelles il appartient. Il est au barycentre de ses modalités.
- la coordonnée d'une modalité j est égale à la moyenne arithmétique des coordonnées des n_j individus de cette modalité. Elle est au barycentre des individus l'ayant pour modalité.

La représentation simultanée des individus et des modalités des variables est donc également possible. En pratique, elle est peu utilisée pour des questions de lisibilité, mais surtout ne revêt que peu d'intérêt la plupart du temps, du fait de « l'anonymat » des individus statistiques (*cf.* ci-dessus partie 1.4), par exemple dans le cas de prélèvements systématiques où l'identité de l'échantillon n'a pas d'importance spécifique. Bien évidemment lorsque les individus ne sont pas anonymes, par exemple dans l'étude des caractéristiques de races de chien (Figure 10) ou dans celle d'établissements archéologiques bien identifiés (*cf.* 2.2.8), la représentation simultanée est intéressante.

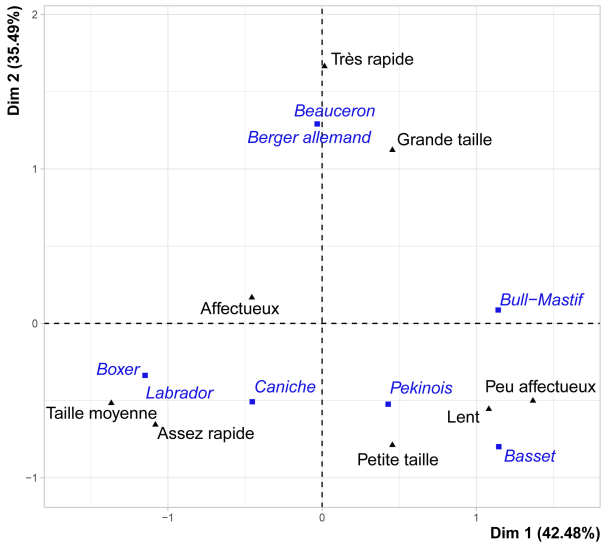


FIG. 10 : Représentation simultanée individus – variables en ACM
Individus : basset, beauceron, berger allemand, boxer, bull-mastif, baniche, labrador, pékinois ; variables : taille (petite, moyenne, grande), vitesse (lent, assez rapide, très rapide), caractère (peu affectueux, affectueux)

3.2 Interprétations statistiques et conseils de mise en œuvre

Interprétations statistiques

La proximité entre individus s'interprète en termes de ressemblance. Deux individus se ressemblent s'ils présentent globalement les mêmes modalités.

Par construction, les modalités d'une même variable s'excluent. Aussi, leur proximité s'interprète en termes de ressemblance, vis-à-vis des autres variables de l'analyse, entre les individus qui présentent ces modalités.

Enfin, la proximité entre des modalités de variables différentes s'interprète en termes d'association. En effet, ces modalités correspondent aux points moyens des individus qui les présentent et sont proches parce qu'elles concernent globalement les mêmes individus ou des individus semblables.

Mise en œuvre

Les propriétés algébriques et géométriques de l'ACM induisent des relations entre l'inertie et les modalités, de même que leur nombre. Ainsi, la part d'inertie due à une modalité est d'autant plus grande que l'effectif de cette modalité est faible. Une modalité sera d'autant plus éloignée du centre que son effectif sera faible. Afin d'éviter de perturber les directions des premiers axes factoriels, il est donc conseillé d'éviter les modalités d'effectif trop faible vis-à-vis des autres. Qui plus est, cela pourrait conduire à une condensation du nuage du reste des modalités autour de l'origine, ce qui permet difficilement de les différencier d'une part et d'observer les relations potentielles qu'elles entretiennent d'autre part[11].

Par ailleurs, la part d'inertie due à une variable est d'autant plus importante que son nombre de modalités est élevé. Il est donc conseillé d'éviter de trop grandes disparités entre le nombre de modalités des p variables sous peine d'artificiellement augmenter l'importance accordée par l'analyse à certaines variables, et de là encore masquer d'éventuelles liaisons existantes avec les autres variables. En effet, les variables ayant peu de modalités, et donc mécaniquement une part d'inertie plus faible, ne prendraient que très peu part à la constitution des axes factoriels. L'étude de leurs liens avec les autres variables sur les plans factoriels ainsi générés serait alors rendue impossible.

3.3 Choix du nombre d'axes factoriels et inertie

Afin de choisir le nombre d'axes factoriels à étudier, le critère du coude précédemment évoqué (*cf.* ci-dessus partie 1.3) reste valide. Le critère de Kaiser consiste quant à lui à retenir les axes factoriels[12] α dont les valeurs propres λ^α sont supérieures à la moyenne des valeurs propres, à savoir $\frac{1}{p}$.

---

[11] *Note des éditeurs* : pour illustrer cet effet, nous pouvons prendre l'exemple de l'analyse du peuplement (*cf.* 2.2.8). Dans ce cas, avec l'ACM nous cherchons à différencier les établissements archéologiques pour les ordonner ensuite en fonction de leur importance dans le système de peuplement. Si, nous incluons dans l'analyse la capitale de cité ou l'*oppidum* principal d'un territoire avec les modalités descriptives spécifiques que seul ou presque cet établissement partagera, nous risquons d'obtenir un plan factoriel (1,2) qui distingue cet individu statistique, ici un établissement particulier, de tous les autres individus (autres établissements) qui se retrouveront dans un nuage de points condensé au centre car finalement, relativement à l'établissement capitale/*oppidum*, ils ont plus de modalités en commun. C'est un résultat qui n'apporte rien puisqu'on connaissait déjà la position de la capitale/*oppidum* par rapport aux autres établissements. Aussi, même si cela peut paraitre contre-intuitif, il vaut mieux supprimer cet établissement très particulier de l'analyse statistique, c'est-à-dire du calcul des axes factoriels, pour observer aisément la manière dont tous les autres établissements du système se positionnent les uns relativement aux autres dans le nuage de points. Il s'agit ensuite de réintroduire cet individu particulier dans l'analyse en l'utilisant comme individu supplémentaire.

[12] Le nombre maximum d'axes factoriels est de $\sum_{i=1}^{p} m_i - p$ où $m_i =$ nb de modalités de la variable i

En ACM, le pourcentage d'inertie dérivé des valeurs propres n'a plus qu'un intérêt restreint. En effet, contrairement à l'ACP et l'AFC, il ne dépend plus directement des liens entre les variables mais seulement du nombre de modalités m en jeu[13]. Il n'a donc plus aucune signification statistique et il n'est plus possible de parler de part d'information restituée par chaque axe factoriel. Benzécri a proposé une correction de ce pourcentage d'inertie afin d'y remédier (Benzécri, 1979). Pour les axes factoriels dont les valeurs propres sont supérieures au seuil de $\frac{1}{p}$, la correction suivante, dépendant du nombre p de variables utilisées dans l'ACM, est calculée :

$$\lambda'^{\alpha} = \left[\left(\frac{p}{p-1}\right)\left(\lambda^{\alpha} - \frac{1}{p}\right)\right]^2 \tag{1.2}$$

La somme S' des valeurs propres corrigées λ'^{α} est alors utilisée pour calculer les pourcentages d'inertie expliquée par chaque axe.

Prenons l'exemple suivant :

Soit $p = 4$ variables qualitatives, les valeurs propres sont corrigées selon Benzécri en utilisant la formule (1.2) comme illustré dans la figure 11.

3 axes ont une valeur propre supérieure au seuil de $\frac{1}{p} = 0{,}25$

		Benzécri		
Axe factoriel	λ^{α}	λ'^{α}	% d'inertie expliquée	% d'inertie cumulée
Axe 1	0,5501	0,1601	54,5	54,5
Axe 2	0,4832	0,0967	32,9	87,4
Axe 3	0,3940	0,0369	12,6	100,0
Axe 4	0,1811			
Axe 5	0,0776			
Axe 6	0,0355			
Axe 7	0,0299			
Somme	1,7514	0,2937		

$\frac{\lambda'^{\alpha}}{S'}$

$S' = \sum_{\alpha} \lambda'^{\alpha}$

FIG. 11 : Correction d'après Benzécri de l'inertie d'une ACM réalisée sur 4 variables

Outre l'interprétation, à nouveau possible, d'inertie expliquée par chaque axe factoriel, cette correction peut permettre d'obtenir un diagramme des valeurs propres plus « parlant », *i.e.* avec un « coude » plus marqué.

3.4 Aides à l'interprétation

Comme en AFC et en ACP, on repèrera les modalités ayant les plus fortes contributions absolues comme les plus intéressantes. On peut retenir le critère suivant :

CTA $> \frac{n_j}{np}$ où n_j = effectif de la modalité j

Ce qui revient à retenir les modalités dont la contribution est supérieure au poids.

[13] De plus, selon que l'on réalise l'ACM sur le tableau disjonctif complet ou sur le tableau de Burt, les valeurs propres obtenues, et donc les pourcentages d'inertie correspondants à chaque axe factoriel, sont différentes.

Pour examiner la contribution d'une variable, on calcule la contribution cumulée qui est égale à la somme des contributions de ses modalités.

L'étude des $\cos^2$ pour la bonne représentation des individus et des modalités est identique aux analyses factorielles précédentes.

3.5 Variables supplémentaires

Il est également possible en ACM d'afficher des variables supplémentaires, ne participant pas à l'analyse et donc au calcul des axes factoriels. Leur représentation dépend de leur nature/type (Figure 12) :

- **Variable qualitative**
  Selon le principe barycentrique, chaque modalité est le point-moyen du groupe d'individus possédant cette modalité, il est ainsi aisé de placer les modalités des variables supplémentaires sur le plan factoriel choisi. Comme en ACP, les valeurs-test des modalités de variables qualitatives supplémentaires sont calculées et interprétées de la même façon (*cf.* ci-dessus partie 1.6).

- **Variable quantitative**
  L'ACM calcule les corrélations de cette variable avec les axes factoriels. On obtient ainsi ses coordonnées dans un cercle de corrélation du type ACP et donc sa représentation dans ce même cercle. Il en résulte un affichage à part et donc un graphique supplémentaire. La position de cette variable sera interprétée en termes de direction dans laquelle se trouvent ses fortes valeurs. Bien sûr ceci est d'autant plus vrai que cette variable est proche du cercle des corrélations, et donc bien représentée. Cette interprétation est à mettre en parallèle avec celles faites sur le même plan factoriel pour les variables qualitatives de l'ACM.

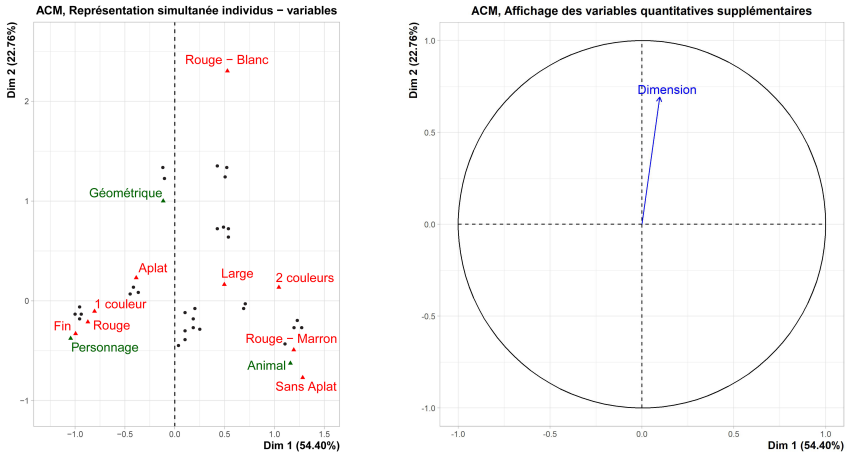


FIG. 12 : ACM des caractéristiques de peintures rupestres relevées sur des parois.
Les individus actifs sont en noir, les variables actives en rouge (épaisseur du tracé, remplissage par aplat, nombre de couleurs utilisées, couleurs), la variable illustrative qualitative est en vert (motif : 3 modalités) et la variable illustrative quantitative en bleu (dimension).

3.6 Intérêts de l'ACM

Cette méthode peut répondre à divers objectifs et donc être employée pour des questionnements différents :

- **Détecter les liaisons entre les variables**
  De même qu'en ACP, les liaisons mises en évidence le sont de façon relative aux liaisons présentes entre les variables étudiées, il est donc nécessaire de revenir aux données, et d'effectuer les tests statistiques adéquats (test du χ^2 d'indépendance ou test exact de Fisher, *cf.* 8), avant toute conclusion.
- **Détecter des individus exceptionnels et/ou d'éventuels groupes d'individus**
  Là encore, se rappeler que cette méthode étant purement descriptive elle ne permet pas sur sa seule base d'effectuer des regroupements d'individus, mais uniquement d'observer et de formuler des hypothèses sur l'existence d'éventuels groupes de caractéristiques similaires. Pour le confirmer, il faudra utiliser des méthodes statistiques adaptées, telles que présentées au chapitre 9.
- **Transformer des variables qualitatives en variables quantitatives** dans le but de pouvoir les utiliser dans des analyses ne s'appliquant qu'à ces dernières.
  Pour cela, ce sont les projections sur les axes factoriels qui seront retenues. Concrètement cela signifie d'utiliser les coordonnées factorielles de chaque individu, sur chaque axe factoriel retenu, comme nouvelles variables (*cf.* 2.2.8).
- **Réduire les dimensions d'un problème**
  De façon similaire à l'ACP, l'utilisation des premiers axes factoriels, au lieu des variables d'origine plus nombreuses, permet de filtrer les fluctuations aléatoires des données et de diminuer le nombre de variables étudiées sans perdre beaucoup d'information. Une ACM peut ainsi être très utile en préalable à une CAH (*cf.* 9 et pour un exemple archéologique 2.2.8).
- **Expliciter les groupes/classes obtenues par une classification**
- **Afficher des variables ou des individus supplémentaires/illustratifs**
- **Permettre de traiter simultanément des variables quantitatives (en les discrétisant) et des variables qualitatives**
  Il est à noter qu'il existe également pour cela des analyses factorielles spécifiques, que nous ne traiterons pas ici, telle que l'Analyse Factorielle de données Mixtes (ADFM), utilisant les variables qualitatives et quantitatives non transformées.

3.7 Bonnes pratiques en ACM

Comme nous l'avons vu au paragraphe 3.2, les liens existants entre l'inertie et l'effectif d'une modalité, et le nombre de modalités elles-mêmes impactant fortement les conclusions de l'ACM, il est préconisé de porter une attention particulière à certains points lors de sa mise en œuvre :

- **Effectuer cette analyse sur un nombre n confortable d'individus** de façon à ce qu'il n'y ait pas trop de modalités par rapport à l'effectif. Toutefois il n'existe pas de seuil théorique auquel se référer, c'est donc l'expérience, la pratique mais aussi le bon sens qui priment en la matière.

- **Eviter les tableaux de contingence trop clairsemés**, *i.e.* présentant beaucoup de valeurs nulles.
- **Être attentif à l'effectif de chaque modalité**
  Empiriquement, on conseille que l'effectif d'une modalité ne soit pas inférieur à $\sqrt{3n}$. Il est donc conseillé de regrouper les modalités rares, tout en maintenant une cohérence, une signification dans la discipline étudiée bien sûr ; et d'éviter les codages trop fins pour les variables quantitatives discrétisées.

Enfin, si les variables sont de nature très différente (paramètres physiologiques, sociologiques, etc…), il sera plus pertinent de procéder à des analyses par sous tableaux ou utiliser des méthodes plus appropriées telle que, par exemple, l'analyse canonique.

En effet, nous avons vu dans ce chapitre les trois analyses factorielles les plus usitées, mais il en existe bien d'autres, spécifiques à certains types de données ou questionnements.

Ouvrages
De nombreux ouvrages traitent des analyses factorielles, en voici deux, faciles d'accès, combinant théorie et cas concrets avec le détail de leurs interprétations :

- Escofier, B. et Pagès, J. (2008). *Analyses factorielles simples et multiples : Objectifs, méthodes et interprétation (4e ed)*. Dunod.
- Husson, F., Pagès, J. et Lê, S. (2009). *Analyse de données avec R*. Presses universitaires de Rennes.

Ressources en ligne
Supports de cours et/ou tutoriels associés :

- *https ://cours-machine-learning.blogspot.com* (R. Rakotomalala, Université Lumière Lyon 2)
- *http ://factominer.free.fr* (F. Husson, S. Lê et J. Josse, Institut Agro Rennes-Angers / INRIA)
  Site support du package FactomineR (Lê et al., 2008), dédié à l'analyse multivariée, détaillant non seulement son utilisation, mais incluant également des rappels théoriques ainsi que des vidéos d'utilisations et d'interprétations d'exemples.

Logiciels
Tous les logiciels de statistiques, SAS, SPSS, Stata, R, etc… sont évidemment à même de réaliser ces analyses. De part son utilisation libre et la quantité de méthodes disponibles, **R** (R Core Team, 2022) et son environnement de développement intégré RStudio ont tendance à s'imposer de plus en plus. Pour les néophytes en programmation, le **package Rcmdr** (Fox, J., 2005) permet de travailler grâce à une interface graphique, et donc par le biais de menus. Le script correspondant à l'action demandée dans le menu étant généré et affiché, cela permet de se l'approprier et de le modifier plus facilement.
Quant aux packages R dédiés aux analyses factorielles, il en existe de nombreux. Personnellement, de part sa simplicité d'utilisation pour les débutants (que ce soit en statistique ou en programmation) et ses nombreuses ressources en ligne, j'ai tendance à conseiller **FactomineR** (plugin disponible pour Rcmdr) et son interface graphique interactive **Factoshiny** (Vaissie, P. et al., 2021).