REPUBLIQUE ALGERIENNE DEMOCRATIQUE et POPULAIRE.
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique.

UNIVERSITE MOULOUD MAMMERI, TIZI-OUZOU
Faculté des Sciences
Département de Mathématiques

Mémoire de Master
en
MATHEMATIQUES

Option
statistique

Thème

Analyse en composantes principales

Présenté par

Hadj kaci karima

Dirigé par:
M^r **Mamou Mohamed.**

Éxaminé par:

Berkoune Youcef	**Maître de conférence**	**UMMTO**	**président**
Mamou Mouhamed	**Maître de Assistant/A**	**UMMTO**	**Rapporteur**
Boualeme Karima	**Maître de Assistant/A**	**UMMTO**	**Examinateur**

Soutenu le 22 / 10 / 2014

Remerciements

Avant de présenter ce travail, mes remerciements vont tout dabord à :

Dieu le tout puissant qui m'a donné la volonté et la santé pour accomplir ce travail et qui m'a aidé à franchir un pas vers le chemin du savoir.

mon promoteur M^r Mamou pour son suivi permanent et sa gentillesse, sa sympathie, sa disponibilité ainsi que pour ses précieux conseils tout au long du projet.

je tiens aussi à remercier mes enseignants pour leur sacrifice et aide précieux jusqu'à l'accomplissement de ce modeste travail.

Enfin, que tous ceux et celles qui, de loin ou de prés nous ont apporté leur aide et soutien trouveront ici, ma reconnaissance et sympatie.

karima.

Dédicaces

Je tiens à dédier ce travail:

A mes très chers parents moussa et zohra pour l'intérêt qu'ils ont porté à mes études et pour leur sacrifice et soutien durant tout mon parcours. Je prie Dieu le tout puissant de les garder en bonne santé et de les récompenser de toutes les peines et sacrifices données aux quels je ne rendrai jamais assez.
À mes frères mohamed, nabil et smail .
À mes soeurs aziza, nawal et soad .
et à toute ma famille sans oublier mes amis(es).

Karima.

Chapitre 1

Introduction

On désigne par statistique descriptive multidimensionnelle l'ensemble des méthodes de la statistique descriptive (ou exploratoire) permettant de traiter simultanément un nombre quelconque de variables.
Les méthodes les plus classiques de la statistique descriptive multidimensionnelle sont les méthodes factorielles.
Les domaines d'utilisation de ces méthodes sont nombreux et diversifiés: biologie, économétrie, médecine, etc . . .
Il existe une multitude de méthodes factorielles permettant de traiter différenes structures de données:
L'analyse en composantes principales pour un tableau de variables quantitatives, l'analyse factorielle des correspondances pour les tables de contingence, l'analyse factorielle multiple pour les variables qualitatives, et l'analyse discriminante pour la prise en compte d'une partition des individus en groupe.

L'origine de ces méthodes remonte au moins à K.Pearson (1901), mais leur pratique n'est devenue courante que depuis l'ère informatique. Elles ont été surtout développées en France dans les années 60, en particulier par Jean-Paul Benzekri qui a beaucoup exploité les aspects géométriques et les représentations graphiques.
Ainsi, à partir d'un tableau rectangulaire de données comportant les observations de p variables quantitatives sur n individus, on peut obtenir des représentations géométriques de ces individus et de ces variables dans un sous espace de faible dimension grâce à l'analyse en composantes principales (ACP).
Le but de mon travail est de mettre en évidence le rôle de l'ACP dans la pratique.

Ce mémoire s'articule autour de trois chapitres principaux:

Le premier chapitre est consacré à quelques définitions.
La matrice des poids, le vecteur moyen, la matrice de covariance, la matrice de corrélation et la matrice d'inertie.

Le deuxième chapitre sera dédié à la présentation de l'ACP Il consiste à réduire le nombre de variables en les résumant par un petit nombre de variables synthétiques c'est a dire trouver un sous espace F_k de dimension faible ($k < p$) tel que l'inertie du nuage projeté sur F_k soit maximale.

Le troisième chapitre est consacré à la mise en oeuvre pratique de l'ACP.

Ce travail se termine par une conclusion générale.

Chapitre 2

Tableaux de données Résumés numériques et les espaces associés

Introduction

Ce chapitre est une introduction à l'analyse en Composantes Principales (ACP). Les connaissances nécessaires pour aborder ce chapitre sont les méthodes de calcul l'espérance, la matrice de variance-covariance, la matrice de corrélation, d'un centre de gravité, la dispersion, et l'inertie pour un nuage de points. Ces outils sont énoncés brièvement au début du chapitre dans la section "Rappel"

2.1 Données multivariées

2.1.1 Le tableau de données

Il a la forme d'une matrice à n lignes et p colonnes

$$X = \begin{bmatrix} x_{11} & x_{12} \cdots x_{1j} \cdots x_{1p} \\ x_{21} & x_{22} \cdots x_{2j} \cdots x_{2p} \\ \cdot & \\ \cdot & \\ \cdot & \\ x_{i1} & x_{i2} \cdots x_{ij} \cdots x_{ip} \\ \cdot & \\ \cdot & \\ \cdot & \\ x_{n1} & x_{n2} \cdots x_{nj} \cdots x_{np} \end{bmatrix}$$

- Le tableau de données $n \times p$ forme un nuage de n points dans un espace à p dimension, et un nuage de p points dans un espace à n dimensions.
L'élément x_{ij} de la matrice X représente la valeur prise par la jéme variable sur le iéme individu.
La iéme ligne de X représente les valeurs prises par toutes les variables sur le iéme individu, on la note par X_i^t et représente l'individu e_i On obtient ainsi n points de $\mathbb{R}^p$.
La jéme colonne de X représente les valeurs prises par tous les individus sur la jéme variable, on la note par X_t^j et représente la variable X_j; on obtient p points de $\mathbb{R}^n$
$X_i^t = (x_{i1},...,x_{ip})$ et $X_t^j = (x_{1j},...,x_{nj})$

Exemple Les données de ce tableau représentent les résultats obtenus par 6 étudiants à un test de statistique(variable 1) et de géologie(variable 2).

$$X = \begin{bmatrix} 11 & 13.5 \\ 12 & 13.5 \\ 13 & 13.5 \\ 14 & 13.5 \\ 15 & 13.5 \\ 16 & 13.5 \end{bmatrix}$$

Autre exemples de tableaux de données
· I = Ensemble de personnes, J = Ensemble de caractères biologiques (taille, poids, rythme cardiaque, capacité thoracique, ...).
· I = Ensemble d'étudiants, J = Ensemble de matières, x_{ij} étant la note obtenue par l'étudiant i dans la matière j.
· I = Ensemble de pays, J = Ensemble de postes de dépenses publiques (éducation, police, culture, etc.), x_{ij} étant la dépense du pays i pour le poste j en 1988.

Lorsque n et p sont grands, ou moyennement grands, le nombre de données np est trés grand.
Comment tirer le plus d'information de ce tableau?
Des techniques d'analyse des données permettent de le faire.

2.1.2 La matrice des poids des individus

On suppose que chaque individu est muni d'un poids p_i tel que $p_i \geq 0$ et$\sum_{i=1}^{n} p_i = 1$. Dans certain cas, il est utile de travailler avec des poids p_i différents (données regroupées...). Ces poids sont regroupés dans la matrice diagonale D_p de taille n:

$$D_p = \begin{bmatrix} p_1 & & & 0 \\ & p_2 & & \\ & & . & \\ & & & . \\ 0 & & & p_n \end{bmatrix}$$

Si les poids sont égaux à $\frac{1}{n}$ on a $D_p = \frac{1}{n} I_n$ où I_n est la matrice identité

2.1.3 Le point moyen ou centre de gravité

Chaque individu e_i sera considéré comme un élément d'un espace vectoreil $\mathbb{R}^p$(espace des individus). L'ensemble des n individus forme un nuage de points de $\mathbb{R}^p$dont le barycentre est le point g défini par:

$$g = \begin{bmatrix} \overline{X1} \\ \overline{X2} \\ . \\ . \\ \overline{Xp} \end{bmatrix}$$

$$\overline{X}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \tag{2.1}$$

Le point **g** est appelé parfois point moyen du nuage.
On a $\mathbf{g} = X' D_p \mathbf{1}_n$ où $\mathbf{1}_n$ désigne le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1 et X' est la matrice transposée de X
Exemple Le vecteur ligne des moyennes arithmétiques pour l'exemple des notes est

$$\overline{X}' = (\frac{11 + ... + 16}{6}, \frac{13.5 + ...13.5}{6}) = (13.3, 13.5)$$

2.1.4 Données centrées et réduites

En ACP on travaillera toujours sur des données centrées ou centrées réduites .
On note par $\dot{X}$ la matrice des données centrées

$$\dot{X} = X - \mathbf{1}_n \mathbf{g}' \tag{2.2}$$

Centrer les données revient à placer l'origine des axes du nuage au centre de gravité g
On note par Z la matrice des données centrées et réduites avec:
$Z_{ij} = \dfrac{x_{ij} - \overline{X}^j}{S_j}$; elle est donnée par la formule :

$$Z = \dot{X} D_{\frac{1}{S}} \tag{2.3}$$

où $D_{\frac{1}{S}}$ est la matrice diagonale des inverses des écarts types

$$D_{\frac{1}{S}} = \begin{bmatrix} \frac{1}{S_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{S_p} \end{bmatrix}$$

2.1.5 Matrice de variance-covariance et de corrélation

La matrice de variance-covariance V est donnée par la formule :

$$V = \dot{X}' D_p \dot{X} = X^t D_p X - gg' \tag{2.4}$$

$$V = \begin{bmatrix} V_{11} & V_{12} & . & . & . & V_{1p} \\ V_{21} & V_{22} & . & . & . & V_{2p} \\ V_{31} & . & V_{33} & . & . & . & V_{3p} \\ . & . & . & V_{44} & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ V_{p1} & . & . & . & . & . & V_{pp} \end{bmatrix}$$

Remarque On a aussi
$X' D_p X = \sum_{i=1}^{n} p_i e_i e_i' \Rightarrow V = \sum_{i=1}^{n} p_i e_i e_i' - gg'$

La matrice de corrélation notée R est définie par:

$$R = D_{\frac{1}{S}} V D_{\frac{1}{S}} = Z' D_p Z \tag{2.5}$$

$$R = \begin{bmatrix} 1 & r_{12} & . & . & . & r_{1p} \\ r_{21} & 1 & . & . & . & r_{2p} \\ r_{31} & . & 1 & . & . & . & r_{3p} \\ . & . & . & 1 & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ r_{p1} & . & . & . & . & . & 1 \end{bmatrix}$$

Remarque: Les matrices V et R sont symétriques.

2.2 L'espace des individus

L'espace $\mathbb{R}^p$ est muni d'une structure euclidienne afin de pouvoir définir des distances entre individus e_i et e_j. On utilisera la formulation générale suivante: la distance entre deux individus e_i et e_j est définie par la forme quadratique:
$d^2(e_i; e_j) = (e_i - e_j)'M(e_i - e_j)$ où M (métrique) est une matrice symétrique($M' = M$) de taille p définie positive
L'espace des individus est donc muni du produit scalaire: $< e_i, e_j >_M = e_i' M e_j$.
Le choix de M dépend de l'utilisateur. En pratique les métriques usuelles en ACP sont:
soit la métrique $M = I$ (matrice identite de rang p)ce qui revient à utiliser le produit scalaire usuel. ou la métrique diagonale des inverses des variances :

$$D_{\frac{1}{S^2}} = \begin{bmatrix} \frac{1}{S_1^2} & & & & 0 \\ & \frac{1}{S_2^2} & & & \\ & & . & & \\ & & & . & \\ & & & & . \\ 0 & & & & \frac{1}{S_p^2} \end{bmatrix}$$

Ce qui revient à diviser chaque caractére par son écart-type l'avantage que la distance entre deux individus ne dépend plus des unités de mesure puisque les nombres $\frac{x_{ij}}{S_j}$ sont sans dimension, ce qui est trés utile lorsque les variables ne s'expriment pas avec la même unité.
Cette métrique donne à chaque caractére la même importance quelle que soit sa dispersion. L'utilisation de la métrique $M = I$ conduirait à privilégier les variables les plus dispersées, pour lesquelles les différences entre individus sont plus fortes et à négliger les différences entre les autres variables. Donc si les données sont homogénes, on utilise la matrice $M = I$

sinon on utilise $M = D_{\frac{1}{S^2}}$

2.3 L'espace des variables

Chaque variable X_i est une liste de n valeurs numériques qui peut être considérée comme un vecteur X_i de l'espace $\mathbb{R}^n$ appelé espace des variables. La métrique utilisée pour le calcul des distances entre variables est la métrique D_p.
Soit les variables X_1,...,X_p centrées. On a les propriétés suivanes:
Le produit scalaire entre deux variables X_k et X_i est

$$< X_k,X_i >= X_k' D_p X_i = S_{ki} \tag{2.6}$$

Le carré de la norme d'une variable est égale à sa variance

$$\| X_k \|^2 = S_k^2 \tag{2.7}$$

et l'écart-type représente donc sa longueur, Le cosinus de l'angle θ_{ki} entre deux variables X_k et X_i est leur coefficient de corrélation linéaire:

$$\cos(\theta_{ki}) = \frac{< X_k,X_i >}{\| X_k \| \| X_i \|} = \frac{S_{ki}}{S_k S_i}. \tag{2.8}$$

2.4 Inertie

On appelle inertie totale du nuage des points E_1,E_2,...,E_n par rapport au centre de gravité g la moyenne des carrés des distances de ces points au centre de gravité
$I_g = \sum_{i=1}^n p_i d_M^2(E_i,g) = \sum_{i=1}^n p_i \|e_i - g\|_M^2 = \sum_{i=1}^n p_i (e_i - g)' M (e_i - g)$
Cette quantité mesure d'une certaine maniére la dispersion globale du nuage autour de g
L'inertie par rapport à un point quelconque H différent de g est définie par:
$I_h = \sum_{i=1}^n p_i d_M^2(e_i,h)$
$I_h = I_g + d_M^2(g,h)$
Si les données sont centrées :
$I_g = I_0 = \sum_{i=1}^n e_i' M e_i$

1. Si $M = I_p$(données centrées mais non réduites) alors:
   $I_g = trace(\sum_{i=1}^n M e_i p_i e_i') = traceM\dot{X}' D_p \dot{X} = traceMV = traceV$

2. Si $M = D_{\frac{1}{S^2}}$ (données centrées et réduites) alors:
   $I_g = traceMV = trace(D_{\frac{1}{S^2}}V) = trace(D_{\frac{1}{S}}VD_{\frac{1}{S^2}}) = traceR = p$

Chapitre 3

Analyse en composantes principales

L'Analyse en composantes principales (ACP) est une méthode de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables liées entre elles (dites "corrélées" en statistique) en nouvelles variables non corrélées les unes des autres. Ces nouvelles variables sont nommées "composantes principales". Elle permet au praticien de réduire le nombre de variables.

3.1 Mise en oeuvre

L'analyse en conposantes principales (ACP) est une méthode de traitement de données multidimensionnelles qui a pour but les deux objectifs suivants

- Visualiser les données.
- Réduire la dimension effective des données.

Géométriquement, les données multidimensionnelles constituent un nuage de points dans $\mathbb{R}^P$.
Si p est supérieure à 3, ce qui est le plus souvent le cas, on ne peut pas visualiser ce nuage. Le seul moyen de visualiser les données est alors de considérer leurs projections sur des droites, des plans ou éventuellement sur des espace de dimension 3. Ainsi, si $U = (U_1,...,U_p) \in \mathbb{R}^P$ est une direction de projection (c'est-à-dire un vecteur de norme un: $\|U\|^2 = U_1^2 + ... + U_p^2 = 1$) les données projetées $(U^T X_1,...,U^T X_n)$ forment un échantillon de dimension 1 que l'on peut visualiser et qui est donc plus facile à interpréter que l'échantillon de départ $(X_1,..,X_n)$ L'idée de base de L'ACP est de cherche la direction $a \in \mathbb{R}^P$, "la plus

interessante", pour laquelle les données projetées seront le plus dispersées possible, c'est-à-dire la direction qui maximise la variance de l'échantillon unidimensionnel (a^TX_1,....,a^TX_n)
$V_a = \frac{1}{n}\sum_{i=1}^{n}(a^TX_i)^2 - (\frac{1}{n}\sum_{i=1}^{n}(a^TX_i))^2$

$= \frac{1}{n}a^T(\sum_{i=1}^{n}X_iX_i^T)a - \frac{1}{n^2}a^T(\sum_{i=1}^{n}X_i\sum_{i=1}^{n}X_i^T)a = a^TVa$
Par conséquent la direction la plus intéressante est une solution de
max a^TVa avec $a \in \mathbb{R}^p et \|a\| = 1$

3.2 Éléments principaux

3.2.1 Axes principaux

Nous devons chercher la droite de $\mathbb{R}^p$ passant par g ($g = 0$ car les données sont centrées) maximisant l'inertie du nuage projeté sur cette droite.

Recherche de premier axe principal

Soit $C_1 = (c_{11},...c_{i1},..c_{n1})$ Le vecteur des coordonnées de la projection orthogonale des individus sur l'axe $U_1 = (u_{11},....,u_{j1},...u_{P1})$
$c_{i1} =< e_i,U_1 >_M = e_i'MU_1$

$$C_1 = \dot{X}U_1$$

I_1 l'inertie du nuage projeté sur cette droite U_1.
$I_1 = \sum_{i=1}^{n} pid^2(e_i,g) = \sum_{i=1}^{n} pic_{i1}c_{i1} = \sum_{i=1}^{n} pic_{i1}^2 = C_1D_pC_1' = Var(C)$
$= U_1\dot{X}D_p\dot{X}'U_1' = U_1'VU_1 = U_1'VU_1$
On choisira U_1 de façon à maximiser cette derniere quantité. Le probleme est donc:

$$\text{maximiser } U_1'VU_1 \text{ avec } U_1'U_1 = 1 \ (M = I_p)$$

Il s'agit d'un probléme classique d'optimisation sous contrainte que l'on peut solutionner par la méthode Lagrange
On forme le Lagrangien
$L = U_1'VU_1 - \lambda(U_1'U_1 - 1)$, $\lambda \neq 0$.
On dérive par rapport à chacune des p composantes du vecteur U_1 ainsi qu par rapport au multiplicateur de Lagrange λ
On obtient

$$2[VU_1 - \lambda U_1] = 0$$

$$U_1'U_1 = 1$$

En simplifiant, on trouve:

$$VU_1 = \lambda U_1$$
$$U_1'U_1 = 1$$

Le premier axe principal est donc forcément le vecteur propre normé U_1 associé à la plus grande valeur propre λ_1 de (V ou R). L'inertie explique par cet axe est égale à λ_1 Le deuxième axe principal est engendré par le vecteur propre normé U_2, orthogonal à U_1 associé à deuxième la valeur propre λ_2 de S ($\lambda_1 \geq \lambda_2$)
Le troisième axe principal est engendré par le vecteur propre normé U_3, orthogonal à U_1 et U_2 associé à la 3^{eme} valeur propre λ_3
etc.....
Le premier plan principal est engendré par U_1 et U_2
Le deuxième plan principal est engendré par U_1 et U_3
Le troisième plan principal est engendré par U_2 et U_3
etc.....
Les axes principaux sont engendrées par les vecteurs propres normés (U'U=1).

Remarque
Les matrices V ou R sont, par construction, symétriques et semi-définie positives.
Propriété de matrice symétrique

1. Les valeurs propres sont réelles.
2. Deux vecteurs propres associés à deux valeurs propres distinctes sont orthogonaux i.e, $U_1'U_2 = 0$..
3. Les valeurs propres d'une matrice d'inertie sont positives ou nulles car elles sont égales à l'inertie expliquées par les axes principaux.

3.2.2 Composantes principales

Les composantes principales sont les variable C_i défines par les axes principaux:

$$C_i = \dot{X}U_i \tag{3.1}$$

. C_i est le vecteur renfermant les coordonnées des projections des individus sur l'axe défini par U_i avec U_i unitaire.

La variance d'une composante principale est égale à la valeur propre λ:

$$V(C_i) = \lambda_i$$

En effet $V(C) = C'D_pC = U'\dot{X}'D_p\dot{X}U = U'VU$ or: $VU = \lambda U$ donc $V(C) = U'\lambda U = \lambda$
Les composantes principales sont elles-même vecteurs propres d'une matrice de taille n. En effet:
$MVu = \lambda u$ s'ecrit $M\dot{X}'D_p\dot{X}u = \lambda u$. En multipliant à gauche par $\dot{X}$ et en remplaçant $\dot{X}u$ par C on obtient, alors, $\dot{X}M\dot{X}'D_pC = \lambda C$, la matrice $\dot{X}M\dot{X}'$ notée W est la matrice dont le terme général w_{ij} est le produit scalaire $< e_i,e_j >= e_i'Me_j$

3.3 Cas usuel. La métrique $D_{\frac{1}{S^2}}$ ou l'ACP sur données centrées-réduites

Le choix de la métrique M est toujours délicat: seul l'utilisateur peut définir correctement la notion de distance entre individus.
Prendre $M = I$ revient à travailler sur la matrice V des variances-covariances, il n'y a pas alors de distinction entre axes principaux et facteurs principaux. Cependant, les résultats obtenus ne sont pas invariants si on change linéairement l'unité de mesure des variables. Les covariances sont multipliées par un facteur K, la variance par un facteur K^2 si on choisit une unité de mesure K fois plus petite pour une variable. Le choix de $M = D_{\frac{1}{S^2}}$ est le plus communément fait et a pour conséquence de rendre les distances entre individus invariantes par transformation linéaire séparée de chaque variable et de s'affranchir des unités de mesure, ce qui est particuliérment intéressant lorsque les variables sont hétérogénes.
En pratique on travaillera donc sur le tableau centré-réduit Z associé à X et on utilisera la métrique $M = I$.
Comme la matrice de variance-covariance des données centrées et réduites est la matrice de corrélation R, les facteurs principaux seront donc les vecteurs propres successifs de matrice R rangés selon l'ordre décroissant des valeurs propres. $Ru = \lambda u$ avec $\|U\|^2 = 1$.
La première composante principale C (et les autres sous la contrainte d'orthgonalité) est la combinaison linéaire des variables centrées et réduites ayant une variance maximale $C = ZU$.

3.4 Qualité des représentations sur les plans principaux

Le but de l'ACP étant d'obtenir une représentation des individus dans un espace de dimension plus faible que p(dim $\mathbb{R}^p$), la question qui se pose alors est: comment apprécie-t-on la perte d'information subie et de savoir combien de facteurs faut-il retenir?
Le critére habituellement utilisé est celui du pourcentage d'inertie totale expliqué.
On mesure la qualité de sous-espace à k dimension (F_k) par:

$$(\frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{I_g}) * 100 = (\frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}) * 100.$$

Ce pourcentage est appellé parfois: le pourcentage (taux d'inertie) expliqué par le sous-espace F_k.
Si par exemble $(\frac{\lambda_1 + \lambda_2}{I_g}) * 100 = 90\%$, on conçoit clairement que le nuage de points est presque aplati sur un sous-espace de dimension deux et qu'une représentation du nuage dans le plan des deux premiers axes principaux sera satisfaisante.

Combien d'axes faut-il retenir?
Le choix des axes retenus est un peu délicat. On peut donner quelques régles :
Régle de Kaiser en ACP normée: on ne s'intéresse qu'aux axes avec une valeur propre superieure à 1 (inertie d'une variable initiale).
Régle du coude: On observe souvent de fortes valeurs propres au départ puis ensuite de faibles valeurs avec un décrochage dans le diagramme.(voir figure 2.1) On retient les axes avant le décrochage.

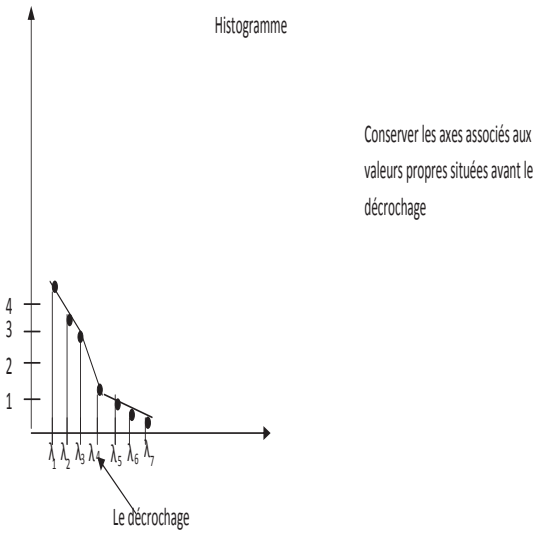

FIG. 3.1

3.5 Les aides à l'interprétation

3.5.1 Corrélations entre composantes principales et variables initiales

La méthode la plus naturelle pour donner une signification à une composante principale C est de la relier aux variables initiales X_j en calculant les coefficients de corrélation linéaire $r(C; X_j)$ en s'intéressant aux plus grandes en valeur absolue .
Lorsque l'on choisit la métrique $D_{\frac{1}{S^2}}$, ce qui revient à travailler sur données centrées-réduites et donc à chercher les valeurs propres et vecteurs propres de R, le calcul de $r(C; X_j)$ est particuliérment simple:
En effet:

$$r(C; X_j) = r(C; Z_j) = \frac{c' D_p Z_j}{S_c}$$

comme $V(C) = \lambda$:

$$r(C; X_j) = r(C; Z_j) = \frac{C' D_p Z_j}{\sqrt{\lambda}}$$

or $C = Zu$ où, facteur principal associé à C, est vecteur propre de R associé à la valeur propre λ:

$$r(C; X_j) = u'Z'D_pZ_j = \frac{(Z_j)'D_pZu}{\sqrt{\lambda}}$$

$(Z_j)'D_pZ$ est la j^{eme} ligne de $Z'D_pZ = R$, donc $(Z_j)'D_pZu$ est la j^{eme} composante de Ru. Comme $Ru = \lambda u$, il vient:

$$r(C; X_j) = \sqrt{\lambda}u_j \tag{3.2}$$

Ces calculs s'effectuent pour chaque composante principale. Pour un couple de composantes principalesC_1 et C_2 par exemple on synthétise usuellement les corrélations sur une figure applée " cercles des corrélations " où chaque variable x_j est repérée par un point d'abscisse $r(C_1; X_j)$ et d'ordonnée $r(C_2; X_j)$. Ainsi la figure montre une premiére composante principalre trés corrélée positivement avec les variables 1, 2 et 3, et négativement avec les variable 4 et 5 et non corrélée avec 6, 7 et 8.

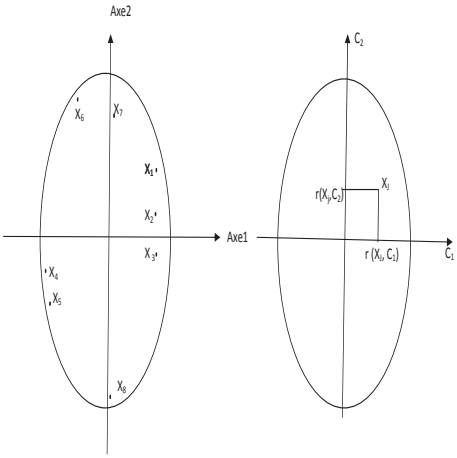

FIG. 3.2

3.5.2 Effet "taille"

Lorsque toutes les variables X_j sont corrélées positivement entre elles, la première composante principale définit un "facteur de taille".
On sait qu'une matrice symétrique ayant tous ses termes positifs admet un premier vecteur propre dont toutes les composantes sont de même signe.

Si on les choisit positives, la première composante principale est alors corrélée positivement avec toutes les variables et les individus sont rangés sur l'axe 1 par valeurs croissantes de l'ensemble des variables (en moyenne). Si de plus les corrélations entre variables sont toutes de même ordre, la premiére composante principale est proportionnelle à la moyenne des variables initiales:

$$\frac{1}{p}\sum_{j=1}^{p} X_j$$

La deuxième composante principale différencie alors des individus de "taille" semblable: on l'appelle facteur de "forme".

3.5.3 La place et l'importance des individus

Il est très utile de calculer pour chaque axe la contribution apportée par les divers individus à la détermination de cet axe. Considérons la k iéme composante C_k; soit c_{ki} la valeur de cette composante pour le i iéme individu. On a : $\sum_{i=1}^{n} p_i c_{ki}^2 = \lambda_k$
La contribution relative de l'individu i à la composante C_k est définie par:

$$CTR_k^{(i)} = \frac{p_i c_{ki}^2}{\lambda_k} \tag{3.3}$$

Cette expression permet de classer les points e_i selon le rôle plus ou moins grand qu'ils ont joué dans la détermination de U_α. Lorsque les poids des individus sont tous égaux à $\frac{1}{n}$, les contributions n'apportent pas plus d'information que les coordonnées

3.5.4 Qualité du positionnement d'un point

Les cosinus carrés sont utilisables pour apprécier la qualité du positionnement des points en représentation factorielle comparé à leur configuration réelle.
En effet, les images obtenues sont des approximations de la configuration réelle.
Il y aura des distances entre couples de points bien représentés, tandis que d'autres ne refléteront pas fidèlement la distance réelle entre les points
Si deux points sont proches du plan factoriel, alors la distance représentée sera une bonne approximation la distance réelle. Mais si au moins un point est éloigné du plan de projection, alors la distance réelle peut être différente de cette proximité du plan factoriel de projection cette distance est mesurée par les cosinus carrés de chaque point avec les axes

factoriels

$$CO2_\alpha(i) = \frac{C_\alpha^2(i)}{d^2(e_i,g)} \tag{3.4}$$

Un cosinus carré égal à 1 indique que l'élément se trouve sur l'axe.
Un cosinus carré égale à 0 indique que l'élément est dans une direction orthogonale à l'axe.

L'addition des cosinus carrés d'un point sur différents axes, donne en pourcentage, la "qualité" de la représentation du point sur le sous-espace défini par ces axes.

3.6 Analyse duale:

Considérons le nuage $N(I)$(nuage de points-individus) et on note par X le tableaux(n, p) des nombres x_{ij}
On a $V = \dot{X}'D_p\dot{X}$ la matrice d'inertie du nuage.
Considérons maintenant le nuage de point-variables $N(J)$ formé de p points X_j situés dans l'espace $\mathbb{R}^n$

$$X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ . \\ . \\ . \\ x_{nj} \end{bmatrix}$$

Définition
On dit que l'analyse de $N(I)$ est l'analyse directe et que l'analyse de N(J) est l'analyse duale de l'analyse de $N(I)$ On a $\Gamma = \dot{X}D_p\dot{X}'$(matrice d'inertie du nuage N(J))
Relation entre les axes factoriels
Pour des raisons de symétrie, les axes factoriels du nuage de points-variables passent par l'origine et ont pour vecteurs directeurs les vecteurs propres unitaires de la matrice Γ.
Si ω_λ est un vecteur propre unitaire de Γ associé à la valeur propre $\lambda_\alpha(\lambda_\alpha \neq 0)$ alors
$U_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\dot{X}'D_p\omega_\alpha$ est un vecteur propre unitaire de V associé à la même valeur propre non nulle λ_α
Inversement si U_α est un vecteur propre unitaire de V associé à la valeur propre non nulle λ_α alors:

$\omega_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \dot{X} U_\alpha D_p$
est un vecteur propre unitaire de Γ associé à la même valeure propre λ_α.

Démonstration

Soit ω_α un vecteur propre unitaire de Γ associé à la même valeur propre λ_α $\lambda_\alpha \neq 0$.
On a: $\Gamma\omega_\alpha = \lambda_\alpha\omega_\alpha \Rightarrow \dot{X} D_p \dot{X}'\omega_\alpha = \lambda_\alpha\omega_\alpha$
En multipliant par $\dot{X}'$ on obtient
$\dot{X}'(\dot{X} D_p \dot{X}'\omega_\alpha) = \dot{X}'(\lambda_\alpha\omega_\alpha) \Rightarrow (\dot{X}' D_p \dot{X})(\dot{X}'\omega_\alpha) = \lambda_\alpha(\dot{X}'\omega_\alpha)$
$\Rightarrow V(\dot{X}'\omega_\alpha) = \lambda_\alpha(\dot{X}'\omega_\alpha)$.
$\Rightarrow \dot{X}'\omega_\alpha$ est un vecteur propre de V associe à la valeur propre λ_α, mais il n'est pas unitaire.
Soit le vecteur de type $k\dot{X}'\omega_\alpha$ unitaire:
$(k\dot{X}'\omega_\alpha)'(k\dot{X}'\omega_\alpha) = 1 \Rightarrow k^2(\omega'_\alpha \dot{X}\dot{X}'\omega_\alpha) = 1$
$\Rightarrow k^2((\omega'_\alpha D_p^{-1}\lambda_\alpha\omega_\alpha) = 1 \Rightarrow k^2\lambda_\alpha D_p^{-1}(\omega'_\alpha\omega_\alpha) = 1$
$\Rightarrow k^2\lambda_\alpha D_p^{-1} = 1 \Rightarrow k = \frac{D_p}{\sqrt{\lambda_\alpha}}$
$\Rightarrow \frac{1}{\sqrt{\lambda_\alpha}} D_p \dot{X}\omega_\alpha$ est un vecteur propre unitaire de V associe à λ_α.
$U_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} D_p \dot{X}\omega_\alpha$ est donc un vecteur propre unitaire de V associe à λ_α

De la même maniére on montre que $\omega_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \dot{X} U_\alpha$
Soit U_α un vecteur propre unitaire de V associe à la valeur propre $\lambda_\alpha \neq 0$
On a: $VU_\alpha = \lambda_\alpha U_\alpha \Rightarrow \dot{X}' D_p \dot{X} U_\alpha = \lambda_\alpha U_\alpha$
On multiplie par $\dot{X}$ on obtient
$\dot{X}(\dot{X}' D_p \dot{X} U_\alpha) = \dot{X}(\lambda_\alpha U_\alpha) \Rightarrow \dot{X} D_p \dot{X}' U_\alpha = \lambda_\alpha(\dot{X} U_\alpha) \Rightarrow \Gamma(\dot{X} U_\alpha) = \lambda_\alpha(\dot{X} U_\alpha)$
$\Rightarrow \dot{X} U_\alpha$ est vecteur propre de Γ associe la valeur propre λ_α, mais il n'est pas unitaire.
Soit le vecteur de type $k\dot{X} U_\alpha$ unitaire:
$(k\dot{X} U_\alpha)'(k\dot{X} U_\alpha) = 1 \Rightarrow k^2(U_\alpha \dot{X}'\dot{X}) = 1$
$\Rightarrow k^2 D_p^{-1}(U'_\alpha \lambda_\alpha U_\alpha) = 1$
$\Rightarrow k = \dfrac{D_p}{\sqrt{\lambda_\alpha}}$

Proposition

Les valeurs propres positives de V et Γ ont le même ordre de multiplicité.

Conséquence

Les sous espaces propres relatifs à même valeur propre $\lambda \neq 0$ ont même dimension.
Ce qui implique que V et Γ ont le même rang.

Coordonnées des points-variables sur les axes

Le vecteur de coordonnées $H_k = (h_{1k},...,h_{pk})$ des variables sur le k^{eme} axe factoreil du nuage de points-variables est donné par

$H_k = \sqrt{\lambda_k} U_k = \dfrac{\dot{X}' D_p C_k}{\sqrt{\lambda_k}}$

$h_{jk} = \sqrt{\lambda_k} U_{jk} = \dfrac{\dot{X}'_j D_p C_k}{\sqrt{\lambda_k}}$

Lorsque l'ACP est normée (X tableau centré réduit), la deuxiéme formule ci-dessus permet de montrer que

$h_{jk} = r(X_j, C_k)$

3.7 Variables et individus supplémentairs

Il peut arriver que l'on veuille représenter sur les différents plans factoriels (c'est à dire plans définis par deux vecteurs propres) des observations ou des variables qui n'étaient pas incluses dans l'analyse initiale. Par exemple, on pourrait disposer d'observations de contrôle ou de nouvelles observations pourraient s'ajouter dans des contextes légèrement différents ou provenant de localisations différentes, . . .Les variables supplémentaires pourraient être des variables de nature très différentes des variables originales.

On distinguera le cas des variables numériques supplémentaires de celui des variables qualitatives supplémentaires.

Individus supplémentaires

On peut représenter un nouvel individu e_s qui n'a pas participé à l'analyse, en le projetant directement sur le sous-espace principal engendré par U_1 et U_2 qui

$< e_s, U_1 >$, $< e_s, U_2 >$

où U_1 et U_2 sont les deux premiers axes principaux.

Variables numériques supplémentaires

La variable numérique supplémentaire X_s peut être placée dans les cercles de corrélation : il suffit de calculer le coefficient de corrélation entre chaque variable supplémentaire et les composantes principales C_1, C_2 , . . .c'est à dire:

$r(C_1; X_s)$, $r(C_2; X_s)$

La variable supplémentaire qualitative

Si la variable à mettre en supplémentaire est qualitative, on ne peut plus effectuer la même transformation.

Dans ce cas, on ramène la variable qualitative ayant m modalités, à m groupes d'individus définis par les modalités de la variable. On traite ensuite ces m groupes d'individus comme des individus supplémentaires. Ce sont les centres de gravité de ces groupes d'individus qui vont être positionnés dans l'espaceR^p.

Toute variable qualitative définit une partition des individus en autant de groupes que la variable possède de modalités.
On peut représenter avec des symboles différents ces groupes d'individus définis par chaque modalité. Pour chaque groupe de points, nous pouvons calculer son point moyen ou centre de gravité (voir la figure)

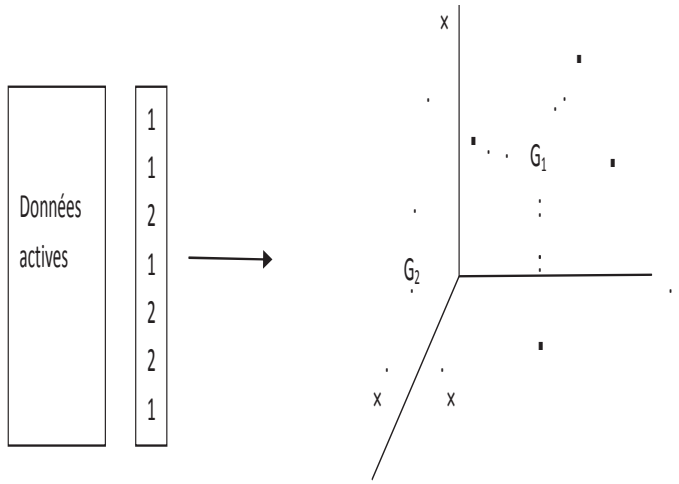

FIG. 3.3

La figure fournit donc une simplification du nuage des points-individus vu du point-de-vue de la variable nominale choisie. La configuration des points-modalités permet en général de qualifier certaines zones du graphique. Elle peut suggérer des éléments d'interprétation des directions factorielles.
Supposons, par exemple, que l'on mesure la taille et le poids de 10 individus et que l'on désire mettre en supplémentaire la variable sexe. Nous disposons des mesures représentées dans le tableau plus bas.

On calcule alors la taille et le poids moyens des hommes $(177; 75)$ et celui des femmes $(167; 59)$. Ce sont ces points moyens qui vont être positionnés parmi les points-individus

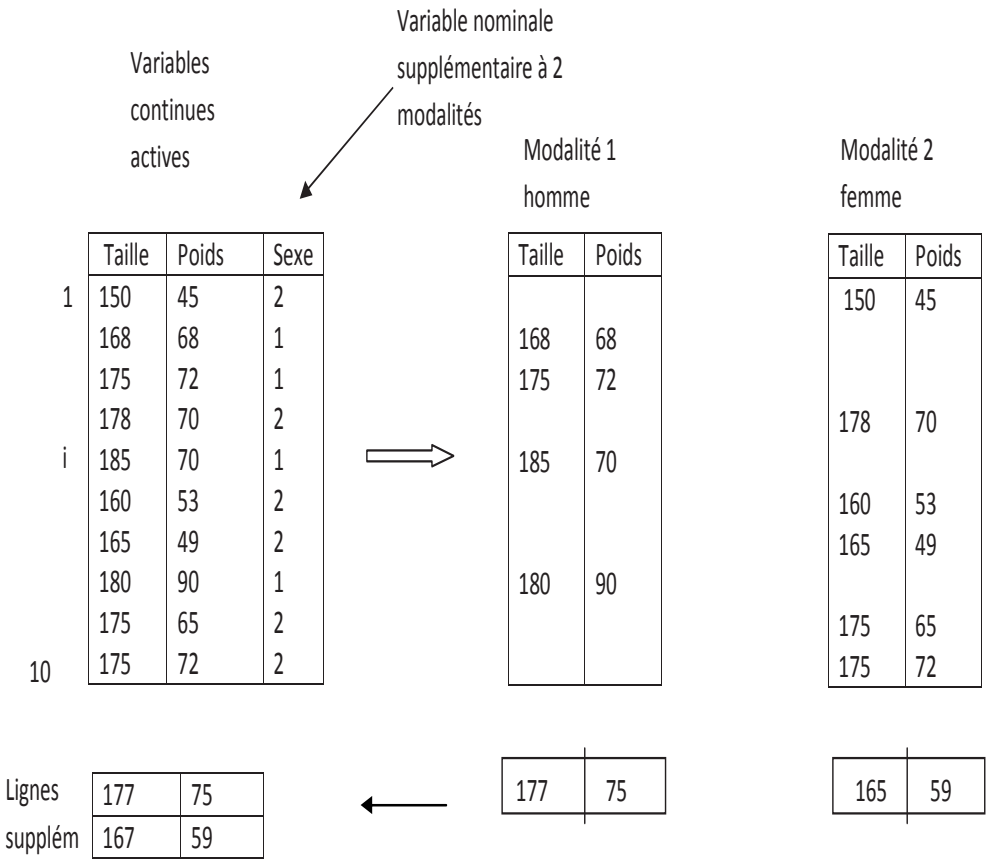

FIG. 3.4

L'analyse d'une variable qualitative supplémentaire ne se fait donc pas dans R^n mais dans R^p La figure schématise le positionnement des modalités supplémentaires comme points moyens des individus qui les composent

3.8 Exemple1

- **ACP centrée**

•Matrice des données X

$$X = \begin{bmatrix} 21 & 38 & 52 \\ 4 & 51 & 67 \\ 67 & 83 & 0 \\ 67 & 3 & 38 \\ 93 & 5 & 6 \end{bmatrix}$$

•Matrice centrée $\dot{X}$

$$\dot{X} = \begin{bmatrix} -29.4 & 2 & 19.4 \\ -46.4 & 15 & 34.4 \\ 16.6 & 47 & -32.6 \\ 16.6 & -33 & 5.4 \\ 42.6 & -31 & -26.6 \end{bmatrix}$$

•Matrice des variance et des covariances $V = (\dot{X}'\dot{X}/n)$

$$V = \begin{bmatrix} 1076.6 & -368.6 & -750.24 \\ -368.6 & 897.6 & -66.2 \\ -750.24 & -66.2 & 671.84 \end{bmatrix}$$

•Valeurs propres de V

$\lambda_1 = 1732.20$

$\lambda_2 = 906.65$

$\lambda_3 = 7.28$

•Vecteurs propres de V

$$U = \begin{bmatrix} u_1 & u_2 & u_3 \\ 0.79 & 0.06 & 0.62 \\ 0.30 & 0.90 & 0.31 \\ 0.54 & 0.43 & 0.73 \end{bmatrix}$$

• **Coordonnées des individus**

$$\begin{bmatrix} & u_1 & u_2 & u_3 \\ obs1 & -34.15 & 8.16 & -3.40 \\ obs2 & -59.54 & 3.84 & 1.03 \\ obs3 & 16.25 & -57.29 & 1.07 \\ obs4 & 20.21 & 31.14 & 3.94 \\ obs5 & 57.24 & 14.14 & -2.65 \end{bmatrix}$$

Ex: observation 4 sur le 2 eme vecteur.
$16.6 * (-0.06) - 33 * (-0.09) + 5.4 * 0.43 = 31.14$

• **Coordonnées des variables**

$$\begin{bmatrix} & \omega_1 & \omega_2 & \omega_3 \\ X_1 & -32.73 & -1.66 & 1.66 \\ X_2 & 12.68 & -27.13 & 0.83 \\ X_3 & 22.37 & 12.95 & 1.96 \end{bmatrix}$$

Ex: variable 2 sur le 1er vecteur:
$-30 * (1732.2)^{\frac{1}{2}} = -12.68$

•**Qualité de représentation des individus**

$$\begin{bmatrix} & u_1 & u_2 & u_3 \\ obs1 & 0.94 & 0.05 & 0.01 \\ obs2 & 1.00 & 0.00 & 0.00 \\ obs3 & 0.07 & 0.93 & 0.00 \\ obs4 & 0.29 & 0.70 & 0.01 \\ obs5 & 0.94 & 0.06 & 0.00 \end{bmatrix}$$

Ex: observation 4 sur le vecteur 2

$$\frac{31.14^2}{(16.6^2 + 33^2 + 5.4^2)} = \frac{969.7}{1393.72} = 0.70$$

•**Qualité de représentation des variables**

$$\begin{bmatrix} & u_2 & u_2 & u_3 \\ X_1 & 0.99 & 0.00 & 0.00 \\ X_2 & 0.18 & 0.82 & 0.00 \\ X_3 & 0.74 & 0.25 & 0.01 \end{bmatrix}$$

Ex: variable 3 sur le vecteur 1
$(-22.37)^2/671.24 = 0.74$

•Contibution des individus

$$\begin{bmatrix} & u_1 & u_2 & u_3 \\ obs1 & 0.13 & 0.01 & 0.32 \\ obs2 & 0.41 & 0.00 & 0.03 \\ obs3 & 0.03 & 0.72 & 0.03 \\ obs4 & 0.05 & 0.21 & 0.43 \\ obs5 & 0.38 & 0.04 & 0.19 \end{bmatrix}$$

Ex: observation 4 sur le 3 eme vecteur

$3.94^2/(5 * 7.28) = 0.43$

• Contribution des variables$(U(.)^2)$

$$\begin{bmatrix} & u_1 & u_2 & u_3 \\ X_1 & 0.62 & 0.00 & 0.38 \\ X_2 & 0.09 & 0.81 & 0.10 \\ X_3 & 0.29 & 0.19 & 0.53 \end{bmatrix}$$

Ex: variable 1 sur le 3eme vecteur
$0.62^2 = 0.38$

3.9 Exemple 2

•ACP centrée réduite

• **Les variables** 1. Var 1: Pain ordinaire.

2. Var 2: Autre pain.
3. Var 3: Vin ordinaire.
4. Var 4: Autre vin.
5. Var 5: Pommes de terre.
6. Var 6: Légumes secs.
7. Var 7: Raisins de table.

Var 8: Plats préparés.

• Les individus

1. Exploitants agricoles.
2. Salariés agricoles.
3. Professions indépendantes.
4. Cadres supérieurs.
5. Cadres moyens.
6. Employés.
7. Ouvriers.
8. Inactifs.

•Matrice des données

$$X = \begin{bmatrix} e_1 & 167 & 1 & 163 & 23 & 41 & 8 & 6 & 6 \\ & 162 & 2 & 141 & 12 & 40 & 12 & 4 & 15 \\ & 119 & 6 & 69 & 56 & 39 & 5 & 13 & 41 \\ & 87 & 11 & 63 & 111 & 27 & 3 & 18 & 39 \\ & 103 & 5 & 68 & 77 & 32 & 4 & 11 & 30 \\ & 111 & 4 & 72 & 66 & 34 & 6 & 10 & 28 \\ & 130 & 3 & 76 & 52 & 43 & 7 & 7 & 16 \\ e_8 & 138 & 7 & 117 & 74 & 53 & 8 & 12 & 20 \end{bmatrix}$$

•Quelques statistique élémentaires

$$\begin{bmatrix} \overline{X} & 127.125 & 4.875 & 96.125 & 58.875 & 38.625 & 6.625 & 10.125 & 24.375 \\ Var & 778.696 & 10.125 & 1504.7 & 980.696 & 61.9821 & 7.98214 & 19.8393 & 149.982 \\ S & 27.905 & 3.182 & 38.790 & 31.316 & 7.873 & 2.825 & 4.454 & 12.247 \end{bmatrix}$$

• **Données centrées réduites**

$$Z = \begin{bmatrix} 1.429 & -1.218 & 1.724 & -1.146 & 0.301 & 0.486 & -0.926 & -1.500 \\ 1.249 & -0.903 & 1.156 & -1.496 & 0.174 & 1.902 & -1.375 & -0.76655 \\ -0.291 & 0.353 & -0.699 & -0.091 & 0.047 & -0.575 & 0.645 & 1.357 \\ -0.437 & 1.924 & -0.853 & 1.664 & -1.476 & -1.283 & 1.768 & 1.194 \\ -0.864 & 0.039 & -O.725 & 0.578 & -0.841 & -0.928 & 0.196 & 0.459 \\ -0.577 & -0.274 & -0.621 & 0.227 & -0.587 & 0.221 & -0.028 & 0.295 \\ 0.103 & -0.589 & -0.518 & -0.219 & 0.555 & 0.132 & -0.701 & -0.683 \\ 0.389 & 0.667 & 0.538 & 0.482 & 1.825 & 0.486 & 0.420 & -0.357 \end{bmatrix}$$

• **Matrice des corrélations multipliées par 100**

$$R = \begin{bmatrix} 100 & -77.366 & 92.619 & -90.579 & 65.635 & 88.856 & -83.343 & -85.585 \\ -77.366 & 100 & -60.401 & 90.444 & -33.285 & -67.337 & 95.882 & 77.122 \\ 92.619 & -60.401 & 100 & -75.016 & 51.708 & 79.173 & -66.901 & -82.799 \\ -90.579 & 90.444 & -75.016 & 100 & -41.857 & -83.860 & 92.393 & 71.979 \\ 65.635 & -33.289 & 51.708 & -41.857 & 100 & 60.292 & -40.993 & -55.396 \\ 88.856 & -67.337 & 79.173 & -83.860 & 60.292 & 100 & -82.445 & -75.092 \\ -83.343 & 95.882 & -66.901 & 92.393 & -40.993 & -82.445 & 100 & 83.445 \\ -85.585 & 77.122 & -82.799 & 71.979 & -55.396 & -75.092 & 83.445 & 100 \end{bmatrix}$$

•Valeurs propres(VP) de la matrice R, et% de la variation comulée totale(VT) expliquée

$$\begin{bmatrix} VP & \%VT \\ 6.21 & 77.6 \\ 0.880 & 88.6 \\ 0.416 & 93.8 \\ 0.306 & 97.6 \\ 0.168 & 99.7 \\ 0.181 & 99.9 \\ 0.00345 & 100 \\ 3.36 * 10^{-12} & 100 \end{bmatrix}$$

Exemple de calcul du % de la variation totale:

$$\% \text{ variation totale } 1 = \frac{6.21}{6.21 + .. + 0.880 + .. + 7.36 * 10^{-10}}$$

•Matrice des vecteurs propres Q de R

$$Q = \begin{bmatrix} -0.391 & -0.138 & 0.162 & 0.119 & 0.294 & 0.398 & -0.107 & 0.728 \\ 0.349 & -0.441 & 0.320 & 0.218 & -0.265 & 0.521 & 0.423 & 0.118 \\ -0.349 & -0.202 & 0.681 & -0.0289 & 0.246 & -0.465 & 0.254 & 0.180 \\ 0.374 & -0.260 & 0.0735 & -0.397 & -0.349 & -0.423 & 0.0333 & 0.578 \\ -0.246 & -0.744 & -0.558 & -0.0740 & 0.176 & -0.108 & 0.0934 & 0.135 \\ -0.365 & -0.128 & 0.0324 & 0.519 & -0.669 & -0.185 & -0.313 & 0.0127 \\ 0.373 & 0.326 & 0.254 & 0.0637 & 0.272 & 0.0163 & -0.766 & 0.159 \\ 0.362 & 0.0502 & -0.162 & 0.708 & 0.333 & -0.360 & 0.225 & 0.219 \end{bmatrix}$$

• Plot des coordonnées des individus sur les deux premières axes principaux (2 premiéres colonnes $Y = RQ$**)**

$$Y = \begin{bmatrix} -3.153 & 0.229 & 0.785 & -0.581 & 0.539 & 0.020 & -0.020 & 3.302 * 10^{-9} \\ -3.294 & 0.418 & 0.328 & 0.857 & -0.641 & 0.004 & 0.029 & -4.575 * 10^{-9} \\ 1.376 & -0.054 & -0.517 & 0.799 & 0.700 & 0.054 & -0.004 & 8.482 * 10^{-10} \\ 4.077 & -0.164 & 0.962 & 0.014 & -0.242 & 0.118 & -0.014 & 3.170 * 10^{-9} \\ 1.607 & 0.801 & -0.163 & -0.385 & 0.037 & -0.130 & 0.109 & -1.829 * 10^{-8} \\ 0754 & 0.756 & -0.322 & -0.064 & -0.192 & -0.183 & -0.102 & 1.556 * 10^{-8} \\ -0.841 & 0.171 & -0.914 & -0.515 & -0.276 & 0.218 & -0.005 & 1.760 * 10^{-9} \\ -0.526 & -2.157 & -0.159 & -0.123 & -0.106 & -0.102 & 0.009 & -1.782 * 10^{-9} \end{bmatrix}$$

On remarque que la variabilité est la plus grand le long de l'axe 1.
Le premier axe met en évidence l'opposition (quant aux consommation étudiées) qui existe entre cadres supérieures (individus 1, 2) et agriculteurs (individu 4). Le deuxième axe est caractéristique des inactifs (individus 8) qui sont opposés à presque toutes les autres catégories.

Interpretion des composantes principales
Ceci se fait en regardant les corrélations avec les variables de départ

$$\begin{bmatrix} Variable & Axe1 : y_1 & Axe2 : y_2 \\ X_1 & \sqrt{\lambda_1}q_{11} = -0.97 & \sqrt{\lambda_2}q_{12} = -0.129 \\ X_2 & \sqrt{\lambda_2}q_{21} = 0.87 & \sqrt{\lambda_2}q_{22} = -0.413 \\ X_3 & \sqrt{\lambda_3}q_{311} = -0.87 & \sqrt{\lambda_2}q_{32} = -0.189 \\ X_4 & \sqrt{\lambda_4}q_{41} = 0.93 & \sqrt{\lambda_2}q_{42} = 0.244 \\ X_5 & \sqrt{\lambda_5}q_{51} = -0.614 & \sqrt{\lambda_2}q_{52} = -0.7 \\ X_6 & \sqrt{\lambda_6}q_{61} = -0.91 & \sqrt{\lambda_2}q_{62} = -0.12 \\ X_7 & \sqrt{\lambda_7}q_{71} = 0.93 & \sqrt{\lambda_2}q_{72} = -0.306 \\ X_8 & \sqrt{\lambda_8}q_{81} = 0.9 & \sqrt{\lambda_2}q_{82} = 0.0471 \end{bmatrix}$$

La première composante principale mesure donc la répartition de la consommation entre aliments ordinaires bon marché (Var1: pain ordinaire, Var3: vin ordinaire, Var6: légumes secs) et aliments plus recherchés (Var2: Autre pain, Var4: Autre vin, Var7: Raisin, Var8:plats préparés) La deuxième composante principale est liée essentiellement à la consommation de pommes de terre dont une consommation élevée caractérise les inactifs.

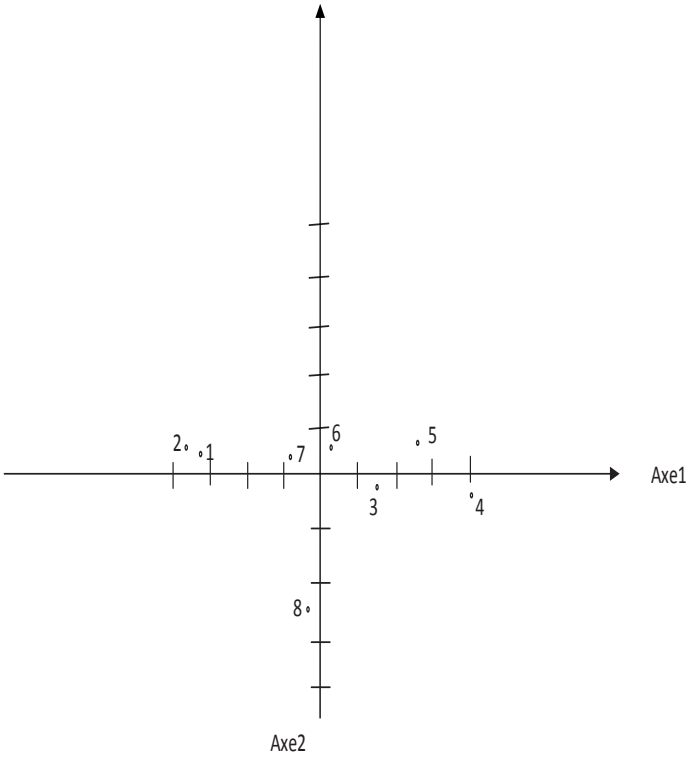

FIG. 3.5 – *Projection des individus sur les deux premier axes principaux*

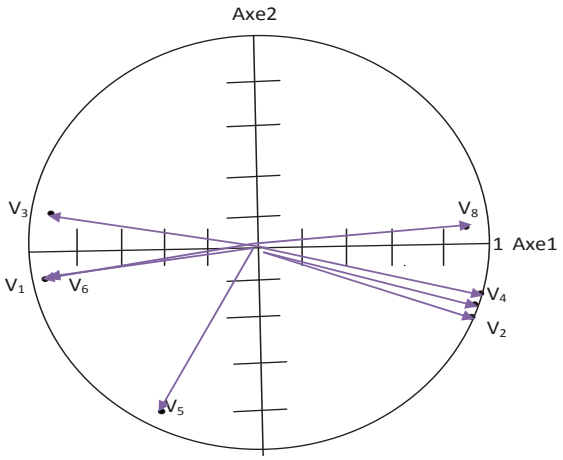

FIG. 3.6 – *Corrélation des données des anciennes variables avec les deux premiers axes principaux.*

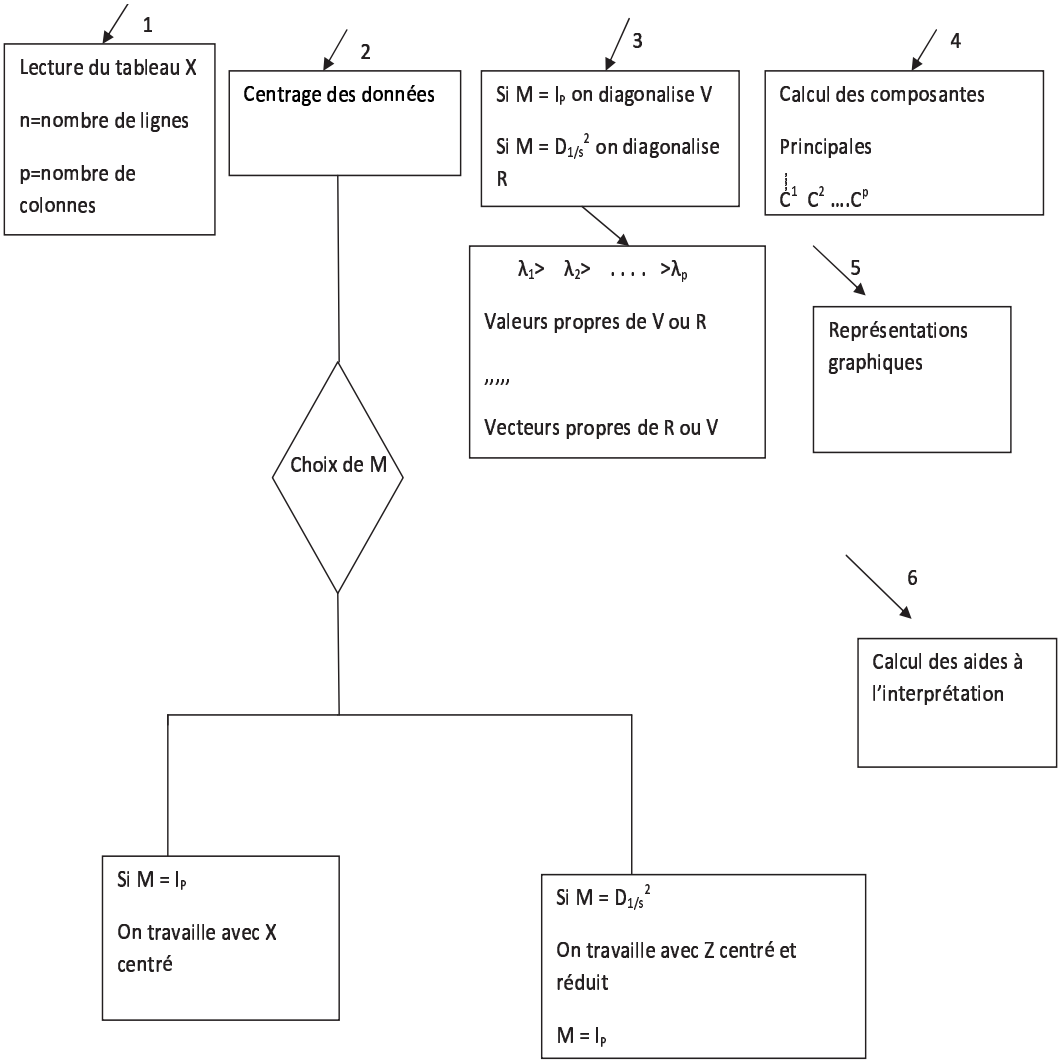

FIG. 3.7 – *résume toutes le étapes pour faire une ACP.*

Chapitre 4

L'application ACP SUR R

Rappels

• L'ACP est une méthode descriptive.

• Son objectif est de représenter sous forme graphique l'essentiel de l'information contenue dans un tableau de données quantitative.

•Dans un tableau de données à j variables, les individus se trouvent dans un espace à j dimensions.

• Lorsqu'on projette ces données sur un plan, on obtient un graphique déformé de la réalité.

• Le rôle de l'ACP est de trouver des espaces de dimensions plus petites minimisant ces déformations.

• On utilise un espace à 2 dimensions (un plan). Ce plan est appelé le plan principal.

Plusieurs packages fournissent des outils permettant de réaliser une analyse en composantes principales. On peut citer :

- dans le package ade4 : dudi.pca

- dans le package FactoMineR : PCA

4.1 Données multivariées

Exemple : Données :

Présentation des données :

Il s'agit d'une enquête (ONU 1967) sur les budgets-temps (temps passé dans différentes activités au cours de la journée).Le tableau suivant comprend 10 variables numériques et 4 variables catégorisées. Les 10 variables numériques sont: le temps passé en: Profession, Transport, Ménage, Enfants, Courses, Toilette, Repas, Sommeil, Télé, Loisirs.

Les 4 variables catégorisées sont: Le sexe (1=Hommes 2=Femmes), l'activité (1=Actifs 2=Non Act. 3=Non précisé), l'état civil (1=Célibataires 2=Mariés 3=Non précisé), le Pays (1=USA 2=Pays de l'Ouest 3=Pays de l'Est 4=Yougoslavie).

Le code suivant est utilisé pour identifier les lignes: H: Hommes, F: Femmes, A: Actifs, N: Non Actifs(ves), M: Mariés, C: Célibataires, U: USA, W: Pays de l'Ouest sauf USA, E : Est sauf Yougoslavie, Y: Yougoslavie

Les temps sont notés en centièmes d'heures. La première case en haut à gauche du tableau (HAU) indique que les Hommes Actifs des USA passent en moyenne 6 heures et 6 minutes (6 heures+10/100 d'heure, soit 6 heures et 6mn) en activité professionnelle. Le total d'une ligne (sur ces 10 variables numériques) est 2400 (24 heures).

Les résultats de l'analyse sont stockés dans la variable budget.pca

library(FactoMineR)

budget.pca < − PCA(budget)

budget.pca$call

$row.w

1	0.03571429	0.03571429	0.03571429	0.03571429	0.03571429	0.03571429
7	0.03571429	0.03571429	0.03571429	0.03571429	0.03571429	0.03571429
13	0.03571429	0.03571429	0.03571429	0.03571429	0.03571429	0.03571429
19	0.03571429	0.03571429	0.03571429	0.03571429	0.03571429	0.03571429
25	0.03571429	0.03571429	0.03571429	0.03571429		

$centre

PROF	*TRAN*	*MENA*	*ENFA*	*COUR*	*TOIL*	*REPA*	*SOMM*	*TELE*	*LOIS*
450.54	86.11	277.07	33.32	108.79	94.93	118.14	785.93	99.43	345.75

$ecart.type

PROF	*TRAN*	*MENA*	*ENFA*	*COUR*	*TOIL*	*REPA*	*SOMM*	*TELE*	*LOIS*
222.72	47.18195.10	29.91	31.97	11.31	25.21	29.23	38.70	62.81	

$X

	PROF	*TRAN*	*MENA*	*ENFA*	*COUR*	*TOIL*	*REPA*	*SOMM*	*TELE*	*LOIS*
HAU	610	140	60	10	120	95	115	760	175	315
FAU	475	90	250	30	140	120	100	775	115	305
FNU	10	0	495	110	170	110	130	785	160	430
HMU	615	140	65	10	115	90	115	765	180	305
FMU	179	29	421	87	161	112	119	776	143	373
HCU	585	115	50	0	150	105	100	760	150	385
FCU	482	94	196	18	141	130	96	775	132	336
HAW	653	100	95	7	57	85	150	808	115	330
FAW	511	70	307	30	80	95	142	816	87	262
FNW	20	7	568	87	112	90	180	843	125	368
HMW	656	97	97	10	52	85	152	808	122	321
FMW	168	22	528	69	102	83	174	824	119	311
HCW	643	105	72	0	62	77	140	813	100	388
FCW	429	34	262	14	92	97	147	849	84	392
HAY	650	140	120	15	85	90	105	760	70	365
FAY	560	105	375	45	90	90	95	745	60	235
FNY	10	10	710	55	145	85	130	815	60	380
HMY	650	145	112	15	85	90	105	760	80	358
FMY	260	52	576	59	116	85	117	775	65	295
HCY	615	125	95	0	115	90	85	760	40	475
FCY	433	89	318	23	112	96	102	774	45	408
HAE	650	142	122	22	76	94	100	764	96	334
FAE	578	106	338	42	106	94	92	752	64	228
FNE	24	8	594	72	158	92	128	840	86	398
HME	652	133	134	22	68	94	102	763	122	310
FME	436	79	433	60	119	90	107	772	73	231
HCE	627	148	68	0	88	92	86	770	58	463
FCE	434	86	297	21	129	102	94	799	58	380

Donc la matrice de données est X

Le vecteur moyen est:

g =(450.54, 86.11, 277.07, 33.32, 108.79, 94.93, 118.14,
785.93, 99.43, 345.75)

Les écarts types des variables sont:

S = (222.72, 47.18, 195.10, 29.91, 31.97, 11.31,

25.21, 29.23, 38.70, 62.81)

La matrice des poids des individus

$$D_p = \begin{bmatrix} 0.03571429 & & & 0 \\ & 0.03571429 & & \\ & & . & \\ & & & . \\ 0 & & & 0.03571429 \end{bmatrix}$$

Description des variables,les histogrammes

•Les histogrammes de toutes les variables

layout(matrix(c(1:10),2,5))

for(i in 1:10) hist(budget[,i],main=names(budget)[i],xlab="")

layout(1)

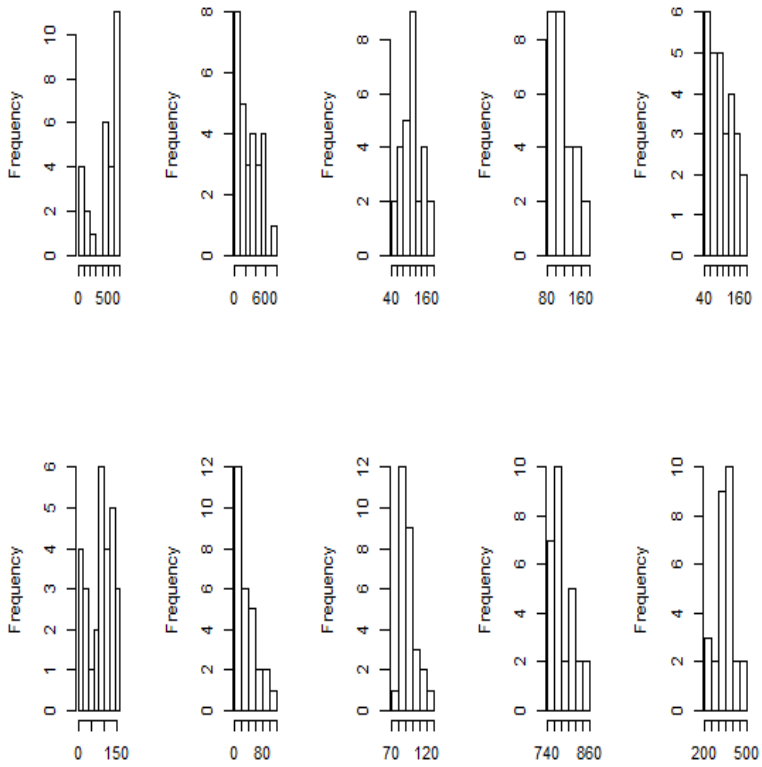

FIG. 4.1

Description des variables, contrôle de la linéarité des relations

•Les relations entre les variables quantitatives

pairs(budget.pca,main="budget-temps")

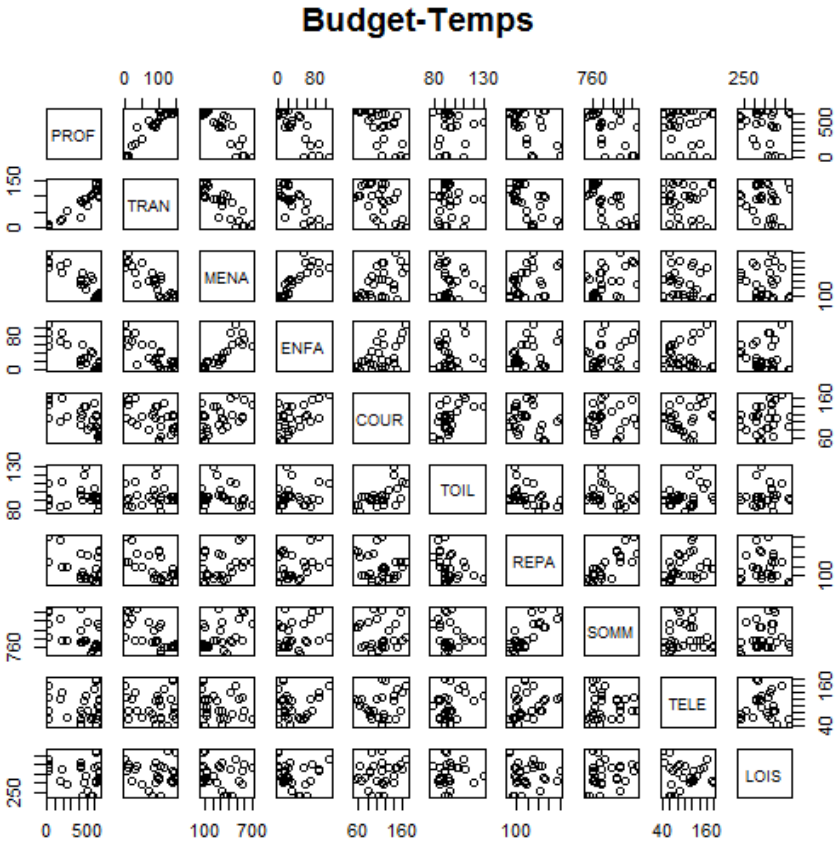

FIG. 4.2

4.2 Analyse en composantes principales

Les résultats relatifs aux valeurs propres

>budget.pc$eig

$$\begin{bmatrix} comp1 & 4.588669e+00 & 4.588669e+01 & 45.88669 \\ comp2 & 2.119843e+00 & 2.119843e01 & 67.08511 \\ comp3 & 1.320978e+00 & 1.320978e+01 & 80.29490 \\ comp4 & 1.195255e+00 & 1.195255e+01 & 92.24745 \\ comp5 & 4.684105e-01 & 4.684105e+00 & 96.93155 \\ comp6 & 1.990474e-01 & 1.990474e+00 & 98.92203 \\ comp7 & 4.681319e-02 & 4.681319e-01 & 99.39016 \\ comp8 & 3.706510e-02 & 3.706510e-01 & 99.76081 \\ comp9 & 2.391893e-02 & 2.391893e-01 & 100.00000 \\ comp10 & 3.472058e-31 & 3.472058e-30 & 100.00000 \end{bmatrix}$$

Les valeurs propres de V sont:
budget.pca$eig[, 1]

λ = (4.588669e+00,2.119843e+00,1.320978e+00,1.195255e+00,4.684105e−01,1.990474e−01,4.681319$e-02$,3.706510$e-02$,2.391893$e-02$,3.472058$e-31$)
Les variances en pourcentages (taux d'inertie) et pourcentage cumulés
budget.pca$eig[, 1]/sum(budget.pca$eig[, 1])*100
on a sum(budget.pca$eig[,1] = I_g = trace de V = 10
4.588669e+01, 2.119843e+01, 1.320978e+01, 1.195255e+01, 4.684105e+00, 1.990474e+00, 4.681319e-01, 3.706510e-01, 2.391893e-01, 3.472058e-30
Donc$\frac{\lambda_1}{I_g}$ = 45.89%

cumsum(budget.pcaeig[,1]/sum(budget.pca$eig[, 1])*100)
45.88669, 67.08511, 80.29490, 92.24745, 96.93155, 98.92203, 99.39016,
99.76081, 100.00000, 100.00000

L'histogramme des valeurs propres
• **Une représentation en % de variance expliquée**
inertie< −budget.pca$eig[,1]/sum(budget.pca$eig[,1]) ∗ 100
barplot(inertie,ylab="% d'inertie", names.arg=round(inertie,2))
title("Eboulis des valeurs propres en %")

Les vecteurs propres associés au cinq premiéres valeurs propres sont données par la matrice U
>budget.pca$svd

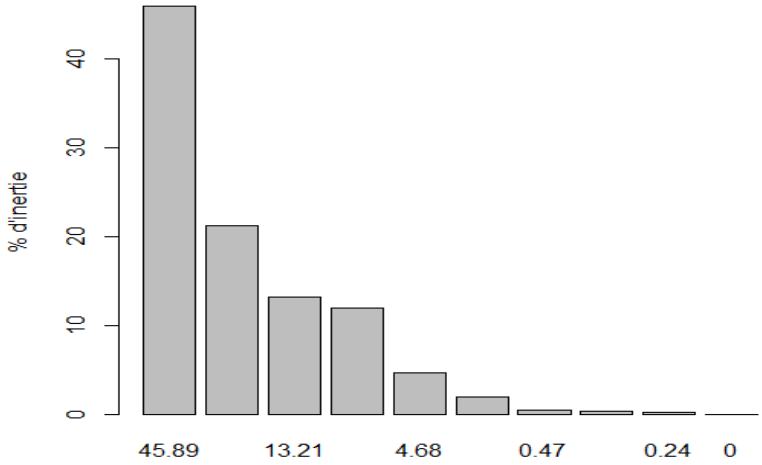

FIG. 4.3 – *Eboulis des valeurs propres en %*

$U

−0.45617162	−0.08313782	0.073587007	−0.06123280	0.140328264
−0.45738827	0.03991548	0.007303258	−0.04166767	−0.162269266
0.42009993	0.01555795	−0.315341555	−0.19589470	0.006074950
0.40712005	0.12264860	−0.072851719	−0.26932467	−0.277549704
0.26310001	0.52241218	0.003967461	0.11071136	−0.124557933
0.03711770	0.56189036	0.262902171	0.05812977	0.655263317
0.27465298	−0.45973933	0.370916117	−0.01293150	0.003644108
0.30072032	−0.39101336	0.166054341	0.28583091	0.457422601
0.04642008	0.13262215	0.809201165	−0.13834203	−0.351950788
0.04303032	0.07572669	−0.026292895	0.87576408	−0.314488720

Les résultats relatifs aux individus

>budget.pca$ind

$coord

FAU	−0.17158846	2.21531946	0.6607635	−0.43755646	1.251658745
FNU	4.05339754	2.27770570	1.0605174	0.52031377	−1.037033431
HMU	−1.77937472	0.29267187	1.8851417	−0.73296793	−1.038424051
FMU	2.61425628	2.28530029	0.7972310	−0.10760356	−0.370677969
HCU	−1.50278808	1.89173274	1.3630367	0.78226738	−0.354485391
FCU	−0.46524227	2.84430152	1.2964113	0.14764754	1.617037527
HAW	−1.17634265	−2.36767945	1.1166316	0.04580158	0.232212327
FAW	0.31200416	−1.49527832	0.2723911	−0.94330781	1.248023435
FNW	4.32338245	−1.63256488	0.8902643	0.14379577	−0.228072827
HMW	−1.12537984	−2.46391614	1.2855676	−0.15025034	0.217813846
FMW	3.13127821	−1.98892449	0.5881848	−0.73455545	−0.345484905
HCW	−1.37003140	−2.57197019	0.5263151	1.02282602	−0.287181199
FCW	1.09911097	−1.65514480	0.5433386	1.49566200	1.429971586
HAY	−2.16269694	−0.24104678	−0.7088986	0.23933281	−0.323902403
FAY	−1.00478037	0.18006790	−1.6157389	−2.13234200	−0.044511065
FNY	3.53743368	−0.37707881	−1.6353265	0.52954143	−0.275953812
HMY	−2.22120121	−0.21162357	−0.4831614	0.10959801	−0.397246578
FMY	1.53997308	−0.21607394	−1.6214171	−1.16576942	−0.439027400
HCY	−2.13527224	0.58061991	−1.6097786	2.17753371	−0.553644702
FCY	−0.33581118	0.41823381	−1.4905524	1.02491288	0.119892949
HAE	−2.14680530	−0.06942418	−0.1312193	−0.32158618	−0.148234546
FAE	−0.98771630	0.58543232	−1.3653640	−2.03994610	0.267280536
FNE	3.91869530	0.04944804	−0.6718550	0.97540533	−0.002319086
HME	−2.07738826	−0.17048157	0.4251206	−0.79232932	−0.216128571
FME	0.49042526	0.20397316	−1.1839185	−2.01251619	0.042539702
HCE	−2.52944831	0.14681468	−1.0625330	1.96343515	−0.351937358
FCE	−0.05514919	0.80353199	−1.0024199	0.96781210	0.842224138

cos2

	$Dim.1$	$Dim.2$	$Dim.3$	$Dim.4$	$Dim.5$
HAU	0.3637026952	0.0544596172	0.405163459	0.0382761019	$8.446371e-02$
FAU	0.0041131789	0.6856053238	0.060994910	0.0267466661	$2.188635e-01$
FNU	0.6577908068	0.2077038840	0.045028239	0.0108387710	$4.305612e-02$
HMU	0.3459667631	0.0093596908	0.388318023	0.0587042433	$1.178281e-01$
FMU	0.5163817724	0.3946036840	0.048022293	0.0008748385	$1.038169e-02$
HCU	0.2456452725	0.3892534615	0.202082186	0.0665615416	$1.366815e-02$
FCU	0.0169664803	0.6341393400	0.131740422	0.0017087790	$2.049621e-01$
HAW	0.1658169048	0.6717488420	0.149410442	0.0002513750	$6.461475e-03$
FAW	0.0199787437	0.4588719043	0.015227667	0.1826224884	$3.196634e-01$
FNW	0.8273763392	0.1179768205	0.035082760	0.0009152674	$2.302517e-03$
HMW	0.1382679108	0.6627885313	0.180431749	0.0024646434	$5.179575e-03$
FMW	0.6598097093	0.2662024081	0.023281072	0.0363098593	8.0321803
HCW	0.1882622087	0.6634893690	0.027783939	0.1049314420	$8.272068e-03$
FCW	0.1399647300	0.3174004297	0.034203983	0.2591805736	$2.369138e-01$
HAY	0.8344618098	0.0103661577	0.089656831	0.0102192641	$1.871731e-02$
FAY	0.1217399564	0.0039098802	0.314798908	0.5482827513	$2.389059e-04$
FNY	0.7554440089	0.0085840040	0.161448721	0.0169287761	$4.597254e-03$
HMY	0.8899089122	0.0080778884	0.042106997	0.0021665817	$2.846362e-02$
FMY	0.3514845970	0.0069196609	0.389645401	0.2014212844	$2.856691e-02$
HCY	0.3597500084	0.0265997867	0.204468595	0.3741313304	$2.418559e-02$
FCY	0.0302385197	0.0469038504	0.595750885	0.2816722829	$3.854406e-03$
HAE	0.8926394472	0.0009334956	0.003334925	0.0200302100	$4.255882e-03$
FAE	0.1304674157	0.0458343432	0.249307027	0.5565126031	$9.553715e-03$
FNE	0.8752373054	0.0001393609	0.025727287	0.0542266876	$3.065324e-07$
HME	0.7426122524	0.0050012854	0.031099325	0.1080284697	$8.038058e-03$
FME	0.0402192023	0.0069571898	0.234385876	0.6772773632	$3.026054e-04$
HCE	0.5430182292	0.0018293709	0.095818118	0.3271869058	$1.051220e-02$
FCE	0.0008598406	0.1825349509	0.284078925	0.2648023227	$2.005373e-01$

contrib

	$Dim.1$	$Dim.2$	$Dim.3$	$Dim.4$	$Dim.5$
HAU	2.446489961	0.792966081	9.46712712	0.988440994	$5.565794e+00$
FAU	0.022915609	8.268202130	1.18042504	0.572070536	$1.194503e+01$
FNU	12.787735654	8.740445649	3.04075785	0.808932006	$8.199754e+00$
HMU	2.464280174	0.144311292	9.60802826	1.605281966	$8.221760e+00$
FMU	5.319264292	8.798829639	1.71836167	0.034596699	$1.047632e+00$
HCU	1.757724217	6.029175383	5.02297915	1.828486782	$9.581026e-01$
FCU	0.168466510	13.629804466	4.54393271	0.065137913	$1.993680e+01$
HAW	1.077017937	9.444612997	3.37105705	0.006268192	$4.111363e-01$
FAW	0.075766292	3.766885607	0.20060061	2.658815731	$1.187574e+01$
FNW	14.547975509	4.490340352	2.14281281	0.061783651	$3.966089e-01$
HMW	0.985719871	10.227987859	4.46823585	0.067454717	$3.617315e-01$
FMW	7.631300634	6.664613199	0.93534938	1.612243204	$9.100672e-01$
HCW	1.460885886	11.144747178	0.74892439	3.125970769	$6.288214e-01$
FCW	0.940239163	4.615410616	0.79815534	6.684182520	$1.559086e+01$
HAY	3.640378833	0.097890610	1.35867167	0.171153531	$7.999147e-01$
FAY	0.785773778	0.054627453	7.05811332	13.586109203	$1.510606e-02$
FNY	9.739392396	0.239553546	7.23028159	0.837878935	$5.806148e-01$
HMY	3.839998475	0.075451252	0.63114738	0.035891094	$1.203194e+00$
FMY	1.845786685	0.078658042	7.10780959	4.060754177	$1.469598e+00$
HCY	3.548638493	0.567965830	7.00613626	14.168084959	$2.337102e+00$
FCY	0.087769892	0.294697630	6.00676656	3.138739551	$1.095980e-01$
HAE	3.587075921	0.008120072	0.04655234	0.309012554	$1.675383e-01$
FAE	0.759310999	0.577419906	5.04014661	12.434226164	$5.446907e-01$
FNE	11.951935567	0.004119424	1.22038662	2.842834827	$4.100614e-05$
HME	3.358849899	0.048965843	0.48861914	1.875827237	$3.561555e-01$
FME	0.187197873	0.070094577	3.78956983	12.102083500	$1.379762e-02$
HCE	4.979742288	0.036314271	3.05232739	11.518992641	$9.443784e-01$
FCE	0.002367192	1.087789097	2.71672447	2.798745949	$5.408424e+00$

Coor (coordonnées des individus sur les cinq premiérs axes principaux $Coor = \dot{X}U$)
Contrib(contribution des individus $CTR_k^i = \frac{p_i C_{ki}^2}{\lambda_k}$)
cos2(qualité du positionnement d'un point individu)

Les résultats relatifs aux variables

> budget.pca$var

coord

	$Dim.1$	$Dim.2$	$Dim.3$	$Dim.4$	$Dim.5$
$PROF$	−0.97717336	−0.12104599	0.084576360	−0.06694442	0.096041398
$TRAN$	−0.97977958	0.05811566	0.008393913	−0.04555431	−0.111057935
$MENA$	0.89990355	0.02265187	−0.362434102	−0.21416721	0.004157728
$ENFA$	0.87209912	0.17857242	−0.083731265	−0.29444652	−0.189956469
$COUR$	0.56359123	0.76061537	0.004559955	0.12103821	−0.085248101
$TOIL$	0.07951049	0.81809434	0.302163513	0.06355195	0.448465641
$REPA$	0.58833905	−0.66936572	0.426308070	−0.01413771	0.002494047
$SOMM$	0.64417836	−0.56930291	0.190852601	0.31249241	0.313062421
$TELE$	0.09943729	0.19309360	0.930045827	−0.15124619	−0.240876960
$LOIS$	0.09217602	0.11025564	−0.030219430	0.95745291	−0.215237725

cor

	$Dim.1$	$Dim.2$	$Dim.3$	$Dim.4$	$Dim.5$
$PROF$	−0.97717336	−0.12104599	0.084576360	−0.06694442	0.096041398
$TRAN$	−0.97977958	0.05811566	0.008393913	−0.04555431	−0.111057935
$MENA$	0.89990355	0.02265187	−0.362434102	−0.21416721	0.004157728
$ENFA$	0.87209912	0.17857242	−0.083731265	−0.29444652	−0.189956469
$COUR$	0.56359123	0.76061537	0.004559955	0.12103821	−0.085248101
$TOIL$	0.07951049	0.81809434	0.302163513	0.06355195	0.448465641
$REPA$	0.58833905	−0.66936572	0.426308070	−0.01413771	0.002494047
$SOMM$	0.64417836	−0.56930291	0.190852601	0.31249241	0.313062421
$TELE$	0.09943729	0.19309360	0.930045827	−0.15124619	−0.240876960
$LOIS$	0.09217602	0.11025564	−0.030219430	0.95745291	−0.215237725

$cos2$

	$Dim.1$	$Dim.2$	$Dim.3$	$Dim.4$	$Dim.5$
$PROF$	0.954867784	0.0146521327	$7.153161e^{-03}$	0.0044815560	$9.223950e^{-03}$
$TRAN$	0.959968034	0.0033774296	$7.045778e^{-05}$	0.0020751954	$1.233386e^{-02}$
$MENA$	0.809826401	0.0005131074	$1.313585e^{-01}$	0.0458675941	$1.728670e^{-05}$
$ENFA$	0.760556867	0.0318881099	$7.010925e^{-03}$	0.0866987526	$3.608346e^{-02}$
$COUR$	0.317635070	0.5785357416	$2.079319e^{-05}$	0.0146502483	$7.267239e^{-03}$
$TOIL$	0.006321918	0.6692783423	$9.130279e^{-02}$	0.0040388503	$2.011214e^{-01}$
$REPA$	0.346142834	0.4480504629	$1.817386e^{-01}$	0.0001998749	$6.220270e^{-06}$
$SOMM$	0.414965757	0.3241058008	$3.642472e^{-02}$	0.0976515093	$9.800808e^{-02}$
$TELE$	0.009887775	0.0372851389	$8.649852e^{-01}$	0.0228754086	$5.802171e^{-02}$
$LOIS$	0.008496418	0.0121563053	$9.132139e^{-04}$	0.9167160684	$4.632728e^{-02}$

contrib

	$Dim.1$	$Dim.2$	$Dim.3$	$Dim.4$	$Dim.5$
$PROF$	20.8092546	0.69118966	0.541504762	0.37494558	1.969202167
$TRAN$	20.9204034	0.15932455	0.005333757	0.17361946	2.633131483
$MENA$	17.6483949	0.02420498	9.944029662	3.83747333	0.003690502
$ENFA$	16.5746732	1.50426783	0.530737297	7.25357756	7.703383798
$COUR$	6.9221615	27.29144840	0.001574075	1.22570059	1.551467857
$TOIL$	0.1377724	31.57207763	6.911755151	0.33790698	42.937001463
$REPA$	7.5434259	21.13602533	13.757876620	0.01672236	0.001327953
$SOMM$	9.0432709	15.28914482	2.757404428	8.16993065	20.923543587
$TELE$	0.2154824	1.75886358	65.480652614	1.91385165	12.386

coor (coordonnées des variables sur les cinq premiérs axes principaux $Coor = U\lambda^{\frac{1}{2}}$)

Contrib(contribution des variables $(U(.)^2)$

cos2(qualité de représentation des variables)

cor(corrélation des variables avec les cinq premiéres cpmposantes principales)

Représentations graphiques

L'objet principal de l'analyse factorielle est de faire figurer des points dans un espace euclidien de faible dimension par rapport à la dimension d'origine. Le but de la représentation graphique est de suggèrer, éclairer, ce que le calcul numérique ne permet pas de saisir. On fera donc des représentations graphiques unidimensionnelles ou bidimensionnelles selon les cas de figures, car on ne peut saisir des représentations de plus de trois dimensions (la dimension 3 pour la visualisation d'un nuage de points n'est pas aisée ; on se contentera des dimensions d'ordre inférieur).

Représentation graphique des lignes (individus)

Les points-lignes (ou points-individus) sont représentés dans l'espace factoriel jugé explicatif. Comme pour les points variables, on procède par projection sur des plans factoriels.

Représentation graphique des variables

Comme les coordonnées factorielles sont assimilables à un coefficient de corrélation, on peut les représenter dans le système d'axes factoriels par rapport à une sphère de rayon unité (un cercle s'il s'agit d'un plan factoriel). On représentera donc successivement les plans factoriels significatifs.

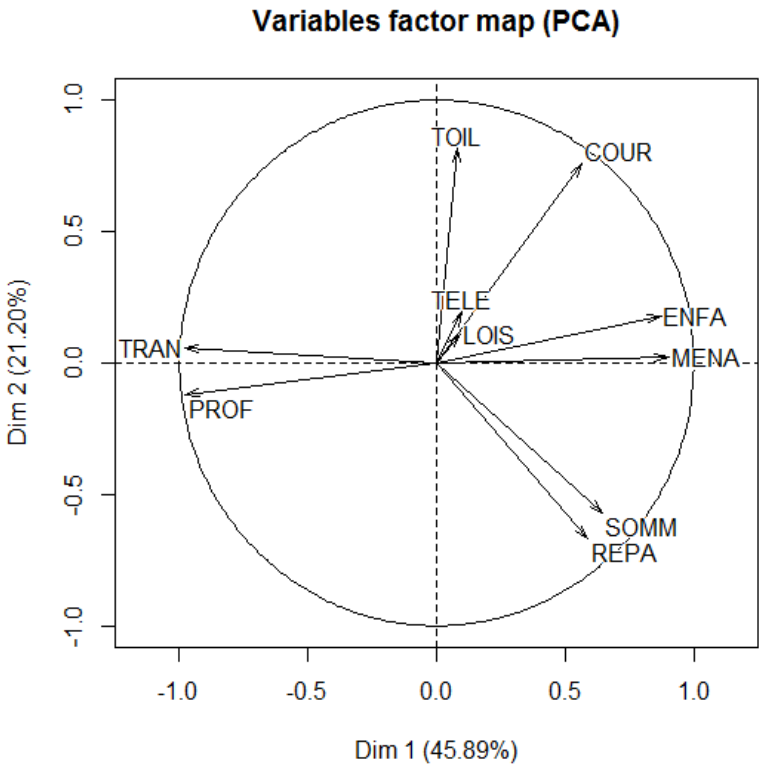

FIG. 4.4

Interprétation

Les variables sont bien représentées dans ce plan factoriel puisque leurs corrélations avec les axes sont relativement importantes (les projections sont proches du cercle de corrélation).
L'interprétation quel'on peut faire des deux premiers axes factoriels est la suivante :
On voit que le premier facteur est correlé positivement, et assez fortement, avec des deux variables: MENE, ENFAN et faiblement avec six variables: SOMM, REPA, COUR, LOIS, TELE, TOIL.
On voit également le premier facteur est corrlé négativement et assez fortement, avec deux variables: TRAN, PROF.
Il s'agit donc d'un axe d'opposition entre les activités d'extérieure et les activités d'intérieurs.
L'axe deux est correlé positivement, et assez fortement, avec deux variables TOIL, TELE, et faiblemene avec quatre variables:COUR, LOIS, ENFAN, TRAN, et nom correlé avec MENA.
D'autre part il correlé négativement avec trois variables: PROF, SOMM, REPA.

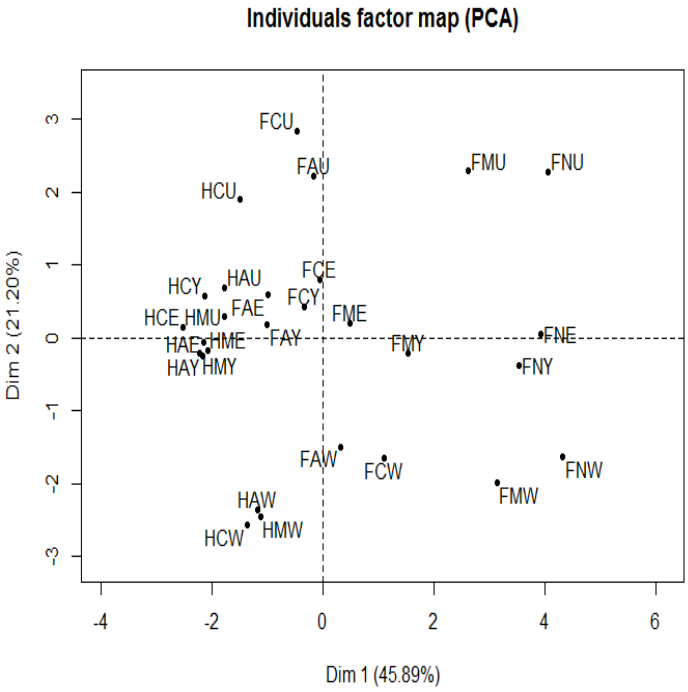

FIG. 4.5

Conclusion

L'ACP est une technique de statistique descriptive dont le principe est simple mais qui met en oeuvre des calculs numériques importants, pour cette raison elle n'a pu se développer qu'avec l'apparition des ordinateurs.

L'ACP est à conseiller pour un premier examen, une mise en forme ou une présentation synthétique de données abondantes croisant des individus avec des variables quantitatives. On n'omettra cependant pas d'examiner préalablement les données par les méthodes statistiques usuelles (moyenne, écart-type, graphiques, corrélation, etc.).

Un reproche fréquemment adressé à l'ACP et aux techniques connexes est qu'elles ne révéleraient que des évidences. Le propos est injuste, mais il est rassurant que souvent les premiers axes retrouvent et confirment ce qui était déjà connu.

Comme avec les autres méthodes descriptives, il faut être très prudent pour inférer des modèles explicatifs ou causals à partir des configurations obtenues.

Bibligraphie

[1] Saporta, G. Probabilités, analyse des données et statistique. Editions Technip, 1990.

[2] Lebart L., Morineau A., Piron M. Statistique Exploratoire Multidimensionnelle.

[3] M. Tenenhaus. Statistique - Méthodes pour décrire, expliquer et prévoir.

[4] J. Pages, B. Escofier Introduction ‘a l’analyse en composantes principales ‘à partir de l’etude d’un tableau de notes.

[5] E.Diday, J.Lemaire, J.Pouget, F.Testu élément d’analyses de données.

[6] Michel Volle professeur à l’EMSAE et au C.E.P.E analyse des données.

[7]Henry Rouanet, Brigitte Leroux analyse des données multimensionnelles statistique en sciences humaines.

[8]Cailliez F.introduction à l’analyse des données.

Table des matières