Faculté des Sciences

Département de Mathématiques

MÉMOIRE DE MASTER

En vue de l'obtention du

Diplôme de Master en Mathématiques.

Option : Probabilité Approfondie et Statistiques

Analyse Multivariée et Applications

Présenté Par : Kasmi Fatima Zohra

Mémoire soutenu le date devant le jury composé de :

M.**A. ALLAM**	Maître de Conférence UABB Tlemcen	Président
M.**A. LABBAS**	Maître de Conférence UABB Tlemcen	Examinateur
M.**S. BOUKHIAR**	Maître Assistant UABB Tlemcen	Encadreur

Année universitaire 2018-2019

Remerciements

C'est grâce à Allah le tout puissant, qui m'a donné la santé, la volonté et la patience que j'ai pu réaliser ce modeste travail.

Je témoigne ici ma profonde gratitude à **Madame Boukhiar Souad** promotrice et directrice de ce mémoire, qui m'a aidé de sa profonde expérience scientifique. Ses conseils pertinents et les nombreuses discussions qu'elle m'a si aimablement accordées malgré ses nombreuses charges ont été un constant encouragement et nul doute qu'ils ont beaucoup contribué à la réalisation de ce modeste travail. Quelle veuille trouver ici l'expression de ma profonde gratitude.

Qu'il me soit permis ici de présenter mes vifs remerciments et mon profond respect à **Monsieur Allam Abdelaziz** qui m'a fait l'honneur de présider le jury de ce mémoire.

J'adresse aussi mes remerciements les plus sincères à **Monsieur Labbas Ahmed** qui a accepté d'examiner ce travail.

Dédicace

A mon très cher père . Toutes les lettres ne sauraient trouver les mots qu'il faut, tous les mots ne sauraient exprimer ma gratitude, mon amour, ma reconnaissance. Rien au monde ne vaut les efforts fournis jours et nuits pour mon education et mon bien être, ce travail est le fruit de tes sacrifices que tu as consentis pour mon education. Tes conseils ont toujours guidé mes pas vers la reussite. Je te dédie ce travail en témoignage de mon profond amour. Qu'Allah, le tout miséricordieux, te préserve, t'accorde santé, bonheur et te protège de tout mal.

A ma très chère mère . Tu représentes pour moi le symbole de la bonté par excellence, la source de tendresse et l'exemple du dévouement qui n'a pas cessé de m'encourager et de prier pour moi. Tes prières et ta bénédiction m'ont été d'un grand secours pour mener à bien mes etudes. Puisse Allah, le tout puissant, te préserver et t'accorder santé, longue vie et bonheur.

A mon très cher frère . Merci de m'avoir accompagné pendant toute ma vie, d'ètre toujours là pour moi et de m'avoir supporté. Qu'Allah le tout puissant, te protège et exauce tous tes voeux.

A ma trés chère grande mère . Qu'Allah l'accueille dans son vaste paradis car si elle etait parmis nous je sais trés bien qu'elle sera fière de sa petite fille.

A toute ma famille .

A mes très chères amies . Bendimerad Sanaa tu es celle qui est toujours là pour moi dans les bons comme dans les mauvais moments, tu es ma moitié, Qu'Allah te garde pour moi Inchaallah, **Boukhalfa Asma, Kada Khadija,Touil wafaa** c'est vraiment une chance pour moi de vous avoir à mes côtés.

A tous les gens que j'aime sans exception.

Table des matières

Introduction

L'analyse des données est un sous domaine des statistiques qui se préoccupe de la description de données conjointes. On cherche par ces méthodes à donner les liens pouvant exister entre les différentes données et à en tirer une information statistique qui permet de décrire de façon plus succincte les principales informations contenues dans ces données. On peut également chercher à les classer en différents sous groupes plus homogènes. Cette méthode consiste essentiellement à établir les relations existants entre les observations, entre les variables ou entre les observations et les variables.

Globalement, ces méthodes sont classées en méthode d'analyse factorielle et méthode des classifications. Elles permettent notamment de manipuler et de synthétiser l'information provenant de tableaux de données d'une taille importante. Pour cela, il est conseillé de bien estimer les corrélations entre les variables que l'on étudie. On a alors souvent recours à la matrice des corrélations.

Le 1^{er} chapitre comporte quelques notions de base en l'algèbre linèaire et aussi quelques résultats importants sur les les vecteurs gaussiens .

Pour le chapitre 2, il s'agit d'une généralisation à k populations de tailles $(n_j)_{1\leq j\leq k}$ des tests de comparaison de moyennes de deux échantillons. Elle permet d'étudier l'effet d'une variable qualitative (facteur) sur une variable quantitative. On utilise des mesures de variance afin de déterminer le caractère significatif, ou non, des différences de moyennes mesurées sur les populations.

Le chapitre 3 est consacré à l'étude de l'analyse des composantes principales "A.C.P" permettant la projection des données de grande dimension dans un espace de dimension plus faible. Nous illustrons cette théorie par un exemple d'application très simple pour visualiser des données réelles.

Dans le chapitre 4, nous étudions l'analyse factorielle des correspondances qui est une A.C.P double anlysant la liaison entre deux variables qualitatives. Plus précisement, elle réalise une A.C.P sur les profils lignes et une autre sur les profils colonnes. Nous expliquons mieu cette technique par l'étude d'un exemple d'application donné en fin de cette partie.

Chapitre 1

Notions de base

1.1 Rappels d'Algèbre Linèaire

1.1.1 Définitions et Notations

Définition 1. ***Espace vectoriel.*** *Un espace vectoriel est une structure stable par addition de vecteurs et par multiplication par un scalaire. Autrement dit , on peut ajouter deux éléments d'un tel espace, ou les multiplier par un nombre , le résultat appartiendra encore à l'espace de départ.*

Soient F et G deux sous espaces vectoriels de E. $F+G=\{x+y\backslash x\in F, y\in G\}$ et $F\times G=\{(x,y)\backslash x\in F, y\in G\}$.

Définition 2. *. On dit que F et G sont supplémentaires si $F\cap G=\varnothing$ et $F+G=E$. De façon èquivalente , tout vecteur x de E s'écrit de manière unique $x=u+v$ avec $u\in F$ et $v\in G$.*

le supplémentaire d'un sous espace véctoriel n'est pas unique.

Définition 3. ***Projecteurs.*** *Si F et G sont supplèmentaires, les applications p et q de E dans E définies par :*

$\forall x \in E$, x=$p(x) + q(x)$ *avec* $p(x) \in F$ *et* $q(x) \in G$ *sont linéaires (on dit que ce sont des endomorphismes de* E *vérifiants) :*

$P1 : p^2 = p, \quad q^2 = q \quad (idempotence)$

$P2 : p \circ q = q \circ p = 0$

$P3 : p + q = I_d$

$P4 : Im(p) = F = ker(q), \quad Im(q) = F = ker(p)$

On dit que p *est la projection sur* F *parallèlement à* G *et que* $q = I_d - p$ *est la projection sur* G *parallèlement à* F.

On appele projecteur dans un espace vectoriel E *tout endomorphisme idempotent de* E.

Dans le cas particulier où les deux sous espaces supplémentaires sont orthogonaux, i.e $E = F \oplus F^{\perp}$ *alors les projecteurs* p *et* q *associées sont dits projecteurs orthogonaux.*

Définition 4. *Une matrice symétrique est une matrice qui est égale à sa propre transposée. Ainsi* A *est symétrique si :* $A^{'} = A$, *ce qui exige que* A *soit une matrice carrée.*

L'ensemble des matrices symétriques à coefficients dans un anneau $\mathcal{K}$ *est noté* $S_n(\mathcal{K})$.

$$\forall S \in S_n(\mathbb{R}), S \in S_n(\mathbb{R})^+ \Leftrightarrow \forall X \in M_{n,1}(\mathbb{R}), X^{'}SX \geq 0.$$

Où $S_n^+(\mathbb{R})$ *est l'ensemble des matrices symétriques positives d'ordre* n.

Définition 5. ***Le rang d'une matrice.*** *Soit* $A \in \mathcal{M}_{n,p}(\mathbb{R})$ *ou* $\mathbb{C}$. *On note* C_1,C_2,...,C_p *les vecteurs colonnes de* A *de* ($\mathbb{R}^n$ *ou* $\mathbb{C}^n$). *On appelle rang de* A *la dimension du sous-espace vectoriel de* ($\mathbb{R}^n$ *ou* $\mathbb{C}^n$) *engendré par* C_1,C_2,...,C_p.

Ou encore, Le rang d'une matrice est le nombre maximum de lignes (ou de colonnes) linéairement indépendantes.

Le rang est toujours inférieur ou égal au minimum du nombre de lignes et du nombre de colonnes de la matrice.

Si le rang de la matrice est égal au minimum du nombre de lignes et du nombre de colonnes, la matrice est dite de plein rang (ou de rang maximal).

Définition 6. ***Matrices inversibles.*** *Une matrice carrée A d'ordre n est dite inversible, s'il existe une matrice B d'ordre n telle que $AB = BA = I_d$, où I_d désigne la matrice unité d'ordre n. Dans ce cas, la matrice B est unique et est appelée la matrice inverse de A, et est notée A^{-1}.*

Si une matrice carrée est de plein rang, alors elle est inversible.

Définition 7. ***Trace d'une matrice.*** *La trace de la matrice carrée A d'ordre n est :*

$$TrA = \sum_{i=1}^{n} a_{ii}$$

Définition 8. ***Valeurs et vecteurs propres.*** *Soit A une matrice $n \times n$, $x \in \mathbb{R}^n$ un vecteur non nul et $\lambda \in \mathbb{R}$. Si $Ax = \lambda x$ Alors (λ, x) sont les éléments propres de A. Il faut que $\det(A - \lambda I_d) = 0$, sinon la seule solution est le vecteur nul.*

Définition 9. *Une matrice carrée A d'ordre n est diagonalisable si elle est de la forme $\Lambda = diag(\lambda_1, ...\lambda_n)$, dans ce cas, il existe une matrice inversible B tel que $\Lambda = B^{-1}AB$.*

La i ème colonne de B est le vecteur propre de A associé à la valeur propre λ_i.

A est diagonalisable si et seulement si ses vecteurs propres sont linéairement indépendants.

1.2 Vecteur gaussiens

Définition 10. *On dit que le vecteur aléatoire $X = (X_1, ...X_n)^{'}$ est gaussien si toute les combinaisons linéaires de X sont gaussiennes. Ou encore, pour tout $u \in \mathbb{R}^n$,*

$$< u, X >= u^{'}X = \sum_{i=1}^{n} u_i X_i$$

Proposition 1. *Si le vecteur aléatoire* $X = (X_1, ..., X_n)'$ *est gaussien alors toute ses composantes sont gaussiennes. La réciprque est fausse.*

Preuve. Soit $X = (X_1, ...X_n)'$ V.G, alors en prenant $u_1 = 1$ et $u_i = 0$ pour tout $i \geq 2$, on en déduit que :

$$X_1 = \sum_{i=1}^{n} u_i X_i$$

est gaussienne. La réciproque est fausse, en effet,

Soit $X \sim \mathcal{N}(0,1)$ et ε une variable aléatoire indépendante de X et suivant la loi de Rademacher : tel que : $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = \frac{1}{2}$. Considérons la nouvelles variable $Y = \varepsilon X$ et le vecteur aléatoire $V = [X, Y]'$. La variable aléatoire Y est gaussienne, en effet :

$$\begin{aligned}
F_Y(u) &= \mathbb{P}(Y \leq u) \\
&= \mathbb{P}((\varepsilon X \leq u) \cap \Omega) \\
&= \mathbb{P}((\varepsilon X \leq u) \cap [(\varepsilon = 1) \cup (\varepsilon = -1)]) \\
&= \mathbb{P}([(\varepsilon X \leq u) \cap (\varepsilon = 1)] \cup [(\varepsilon X \leq u) \cap (\varepsilon = -1)]) \\
&= \mathbb{P}((\varepsilon X \leq u) \cap (\varepsilon = 1)) + \mathbb{P}((\varepsilon X \leq u) \cap (\varepsilon = -1)) \\
&= \mathbb{P}(X \leq u, \varepsilon = 1) + \mathbb{P}(X \geq -u, \varepsilon = -1) \\
&= \frac{1}{2}\mathbb{P}(X \leq u) + \frac{1}{2}\mathbb{P}(X \geq -u) \\
&= \mathbb{P}(X \leq u) = F_X(u)
\end{aligned}$$

Par suite, Y suit la loi normale $\mathcal{N}(0,1)$, par contre, le vecteur $V = [X, Y]'$ n'est pas gaussien, en effet, pour $Z = X + Y = (1 + \varepsilon)X$, on a,

$$\mathbb{P}(Z = 0) = \mathbb{P}(1 + \varepsilon = 0) = \mathbb{P}(\varepsilon = -1) = \frac{1}{2}$$

ce qui est impossible pour une variable gaussienne.

Proposition 2. *Soit le vecteur $X = (X_1, ..., X_n)$ de v.a indépendantes. Il est gaussien si et seulement si pour tout $i \in \mathbb{N}$, X_i est gaussienne.*

Proposition 3. *La matrice de covariance d'un vecteur aléatoire, si elle existe, est symétrique réelle positive, i.e $\forall u \in \mathbb{R}^n \quad u'\Gamma u \geq 0$, elle est donc diagonalisable en base orthonormée :*

$$\Gamma = P'\Delta P,$$

avec $P' = P^{-1}$ et $\Delta = diag\{\lambda_1, ..., \lambda_n\}$, les λ_i étant tous positifs ou nuls.

preuve. Notons par Γ_X la matrice de covariance du V.a X, sachant que $Cov(X,Y)$=$Cov(Y,X)$ (ou comme $(AA')'$=AA'), Γ_X est symétrique. Reste à montrer que pour tout vecteur réel $u = (u_1, ..., u_n)'$, on a $u'\Gamma u \geq 0$. Or

$$u'\mathbb{E}[(X-\mathbb{E}[X])(X-\mathbb{E}[X])']u = \mathbb{E}[(u'(X-\mathbb{E}[X]))((X-\mathbb{E}[X])'u)] = \mathbb{E}[(u'(X-\mathbb{E}[X]))^2] \geq 0.$$

Proposition 4. *Soit X=$(X_1, ..., X_n)'$ un V.a de matrice de covariance Γ_X. La variable aléatoire Z=$\alpha_1 X_1 + ... + \alpha_n X_n$=$\alpha' X$ a pour variance :*

$$Var(Z) = \alpha'\Gamma_X\alpha = \begin{bmatrix}\alpha_1 & . & . & . & \alpha_n\end{bmatrix}\Gamma_X\begin{bmatrix}\alpha_1 \\ . \\ . \\ . \\ \alpha_n\end{bmatrix}$$

Définition 11. ***Fonction caractéristique d'un vecteur gaussien .*** *Soit X un V.a dans $\mathbb{R}^n$, sa fonction caractéristique est donnée par :*

$$\Phi_X : \begin{cases} \mathbb{R}^n \to \mathbb{C}, \\ u = [u_1, ..., u_n]' \to \Phi_X(u) = \mathbb{E}[\exp(i < u, X >)] = \mathbb{E}[\exp(i\sum_{j=1}^n u_j X_j)] \end{cases}$$

Ou encore,

$$\Phi_X(u) = \exp(iu'm - \frac{1}{2}u'\Gamma_X u)$$

Proposition 5. ***Transormation affine .*** *Si $X \sim \mathcal{N}(m, \Gamma_X)$, pour $A \in M_{k,n}(\mathbb{R})$ et $B \in M_{k,1}(\mathbb{R})$, alors le vecteur Y=$AX + B$ est gaussien avec*

$$Y \sim \mathcal{N}(Am + B, A\Gamma A').$$

Définition 12. *Soit X et Y deux variables aléatoires de carrés intégrables et indépendantes alors on dit qu'elles sont non corrélées si $Cov(X,Y)$=0 ou encore la matrice de covariance du vecteur $[X,Y]'$ est diagonale.*

Remarque 1. *L'exemple suivant nous montre que la réciproque est en général est fausse.*

Exemple. Décorrélation $\not\Rightarrow$Indépendance

Soit $X \sim \mathcal{N}(0,1)$ et $Y = X^2$, donc $\mathbb{E}[Y] = \mathbb{E}[X^2] = Var(X) = 1$. En calculant la covariance :

$$Cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$$

donc X et Y sont bien décorrélées. Cependant X et Y ne sont pas indépendantes car Y est une fonction déterministe de X.

Proposition 6. *Indépendance $\Leftrightarrow$ Décorrélation*

Soit X=$(X_1,...,X_n)'$ un vecteur aléatoire gaussien. Les variables aléatoires $(X_1,...,X_n)$ sont indépendantes si et seulement si elles sont non corrélées, c'est-à-dire la matrice de covariance Γ_X est diagonale.

preuve. Supposons X gaussien et de composantes indépendantes. Alors ces composantes sont non corrélées, c'est-à-dire :

$$\forall (i,j) \in \{1,...,n\}^2 \quad Cov(X_i, X_j) = 0,$$

et la matrice Γ_X est digonale. Ceci est toujours vrai, l'aspect gaussien de X n'est pas nécessaire.

Réciproquement, supposons X gaussien et de matrice de covariance Γ_X diagonale

$$\Gamma_X = diag(\sigma_1^2, ..., \sigma_n^2).$$

Si on note $m = [m_1, ..., m_n]^{'}$ la moyenne de X, celui-ci admet pour fonction caractéristique :

$$\Phi_X(u) = \exp(iu^{'}m - \frac{1}{2}u^{'}\Gamma_X u)$$

où

$$\begin{aligned} \forall j \in \{1, ..., n\}, \quad \Phi_X(u) &= \exp(i\sum_{j=1}^{n} m_j u_j - \frac{1}{2}\sum_{j=1}^{n} u_j^t \Gamma_{X_j} u_j) \\ &= \prod_{j=1}^{n} \exp(im_j u_j - \frac{\sigma_j^2 u_j^2}{2}) = \prod_{j=1}^{n} \Phi_{X_j}(u_j) \end{aligned}$$

Proposition 7. *Soit $X = (X_1, ..., X_n)^{'}$ un vecteur gaussien de moyenne m et de matrice de covariance Γ_X. Il existe P orthogonale telle que $P\Gamma_X P^{'} = \Delta = diag(\lambda_1, ..., \lambda_n)$, avec les $\lambda_j \geq 0$. Alors les composantes Y_j du vecteur aléatoire $Y = P(X - m)$ sont des variables aléatoires gaussiennes indépendantes centrées de variances respectives λ_j.*

preuve . Puisque Γ_X est symétrique réelle positive, elle est diagonalisable en base orthonormée : $\Gamma_X = P^{'}\Delta P$, avec :

$$\Delta = diag(\lambda_1, ..., \lambda_n),$$

où les λ_j sont les valeurs propres positives de Γ_X et P une matrice orthogonale. Si on considère maintenant le nouveau vecteur aléatoire

$$Y = (Y_1, ..., Y_n)^{'} = P(X - m) = PX - Pm,$$

c'est encore un vecteur gaussien, en tant que transformée affine d'un vecteur gaussien. Plus précisément, on sait que :

$$Y \sim \mathcal{N}(Pm - Pm, P\Gamma_X P^{'}) = \mathcal{N}(0, \Delta).$$

Ainsi le vecteur gaussien Y est centré et ses composantes sont indépendantes, puisque sa matrice de covariance est diagonale.

Proposition 8. ***Densité d'un vecteur gaussien .***

Si $X \sim \mathcal{N}(m, \Gamma_X)$, avec Γ_X inversible, alors X admet pour densité :

$$f(x) = f(x_1, ..., x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det \Gamma_X}} \exp(-\frac{1}{2}(x-m)^{'}\Gamma_X^{-1}(x-m))$$

preuve . On utilise la transormation affine du résultat précédent : $Y = P(X - m)$, avec :

$$P\Gamma_X P^{'} = \Delta = diag(\lambda_1, ..., \lambda_n).$$

dire que Γ_X est inversible équivaut à dire que les valeurs propres λ_j sont toutes strictement positives. Les composantes $Y_1, ..., Y_j$ sont indépendantes, avec $Y_j \sim \mathcal{N}(0, \lambda_j)$, donc Y admet pour densité :

$$f_Y(y) = \prod_{j=1}^{n} f_j(y_j) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\lambda_j}} \exp(-\frac{y_j^2}{2\lambda_j})$$

qu'on peut encore écrire :

$$f_Y(y) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det \Gamma_X}} \exp(-\frac{1}{2}y^{'}\Delta^{-1}y)$$

Pour retrouver la densité de X, il suffit alors d'appliquer la formule de changement de variable pour le $\mathcal{C}^1$-difféomorphisme

$$\phi : \begin{cases} \mathbb{R}^n \to \mathbb{R}^n, \\ x \to y = P(x-m), \end{cases}$$

Ce qui donne

$$f_X(x) = f_Y(P(x-m))|\det J_\phi(x)|.$$

Or ϕ est une transformation affine, donc $\forall x \in \mathbb{R}^n$, $J_\phi(x)$=P et puisque P est orthogonale alors, $\forall x \in \mathbb{R}^n$, $|\det J_\phi(x)|$=1. On en déduit la densité du vecteur X

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det \Gamma_X}} \exp(-\frac{1}{2}(x-m)^{'}\Gamma_X^{-1}(x-m))$$

1.3 Conditionnement des vecteurs gaussiens

Théoreme. *Soit la v.a Y telle que $\mathbb{E}[Y^2] < +\infty$. Parmi toutes les fonctions $u : \mathbb{R} \to \mathbb{R}$, l'erreur d'aproximation $\mathbb{E}[(Y - u(X))^2]$ est minimale lorsque u est la fonction de régression $\mathbb{E}[Y|X]$.*

preuve . Notons $m(X) = \mathbb{E}[Y|X]$, alors pour toute fonction $u : \mathbb{R} \to \mathbb{R}$, on peut écrire

$$\mathbb{E}[(Y - u(X))^2] = \mathbb{E}[((Y - m(X)) + (m(X) - u(X)))^2]$$

On utilise la linéarité de l'espérance

$$\mathbb{E}[(Y-u(X))^2] = \mathbb{E}[(Y-m(X))^2]+2\mathbb{E}[(Y-m(X))(m(X)-u(X))]+\mathbb{E}[(u(X)-m(X))^2].$$

Or le calcul d'espérance par conditionnement assure que

$$\mathbb{E}[(Y - m(X))(m(X) - u(X))] = \mathbb{E}[\mathbb{E}[(Y - m(X))(m(X) - u(X))|X]],$$

et puisque $m(X) - u(X)$ est une fonction de X, on sait que

$$\mathbb{E}[(Y - m(X))(m(X) - u(X))] = \mathbb{E}[\mathbb{E}[Y - m(X))|X](m(X) - u(X))],$$

Or par linéarité de l'espérance conditionnelle et puisque $\mathbb{E}[m(X)|X] = m(X) = \mathbb{E}[Y|X]$, on en déduit que

$$\mathbb{E}[(Y - m(X))|X] = \mathbb{E}[Y|X] - \mathbb{E}[m(X)|X] = \mathbb{E}[Y|X] - m(X) = 0$$

on a donc obtenu

$$\mathbb{E}[(Y - u(X))^2] = \mathbb{E}[(Y - m(X))^2] + \mathbb{E}[(u(X) - m(X))^2].$$

cette quantité est minimale lorsque $u(X) = \mathbb{E}[Y|X]$. On a vu que le mieux à faire est de prendre u la fonction de régression de Y sur X, c'est-à-dire la fonction qui à X associe

$\mathbb{E}[Y|X=x]$. D'après le théorème de projection, la variable aléatoire $\mathbb{E}[Y|X]$ est la fonction $u(X)$ caractérisée par la double propriété

$$\begin{cases} u(X) \in \mathbf{L}^2(X), \\ Y - u(X) \perp \mathbf{L}^2(X), \end{cases}$$

où $\mathbf{L}^2(X) = \{u(X) \text{ avec } u : \mathbb{R} \to \mathbb{R} \text{ borélienne telle que } \mathbb{E}[u^2(X)] < +\infty\}$

Proposition 9. *La droite de régression est donnée par $f(X) = aX + b$, avec :*

$$\begin{cases} a = \dfrac{Cov(X,Y)}{Var(X)}, \\ b = \mathbb{E}[Y] - a\mathbb{E}[X], \end{cases}$$

c'est-à-dire

$$f(X) = \mathbb{E}[Y] + \frac{Cov(X,Y)}{Var(X)}(X - \mathbb{E}[X])$$

preuve . On cherche la meilleure droite $f(X) = aX + b$ qui approxime Y au sens de l'érreur quadratique $\mathbb{E}(Y - (aX+b))^2$.

$$\begin{aligned} \mathbb{E}(Y-(aX+b))^2 &= \mathbb{E}[(Y - \mathbb{E}(Y) + \mathbb{E}(Y) - a(X - \mathbb{E}(X) + \mathbb{E}(X)) - b)]^2 \\ &= \mathbb{E}[(Y - \mathbb{E}(Y)) - a(X - \mathbb{E}(X)) + (\mathbb{E}(Y) - a\mathbb{E}(X) - b)]^2 \\ &= \sigma_Y^2 + a^2\sigma_X^2 - 2aCov(X,Y) + (\mathbb{E}(Y) - a\mathbb{E}(X) - b)^2 \\ &= \sigma_Y^2 + \left(a\sigma_X - \frac{Cov(X,Y)}{\sigma_X}\right)^2 - \frac{Cov^2(X,Y)}{\sigma_X^2} + (\mathbb{E}(Y) - a\mathbb{E}(X) - b)^2 \\ &= \left(a\sigma_X - \frac{Cov(X,Y)}{\sigma_X}\right)^2 + \sigma_Y^2\left(1 - \left(\frac{Cov(X,Y)}{\sigma_X\sigma_Y}\right)^2\right) + (\mathbb{E}(Y) - a\mathbb{E}(X) - b)^2 \end{aligned}$$

En notant ρ le coefficient de corrélation linéaire

$$\rho = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}.$$

Après dérivation par rapport à a et b, nous trouvons

$$a = \frac{Cov(X,Y)}{\sigma_X^2} \quad et \quad b = \mathbb{E}(Y) - a\mathbb{E}(X)$$

Théoreme. ***Espérance conditionnelle ⇔ droite de régression .***

Si $[X, Y]'$ *est un vecteur gaussien, alors*

$$\mathbb{E}[Y|X] = \mathbb{E}[Y] + \frac{Cov(X,Y)}{Var(X)}(X - \mathbb{E}[X]).$$

preuve . Il suffit de prouver que la fonction u définie par :

$$u(X) = \mathbb{E}[Y] + \frac{Cov(X,Y)}{Var(X)}(X - \mathbb{E}[X])$$

vérifie bien la double propriété de caractérisation de l'éspérance conditionnelle. Puisque X est gaussienne, elle est dans $\mathbf{L}^2(\Omega)$, et par suite $u(X) = aX + b$ est dans $\mathbf{L}^2(X)$. Il reste à prouver que la variable aléatoire $(Y - u(X))$ est orthogonale au sous-espace $\mathbf{L}^2(X)$, c'est-à-dire orthogonale à toute variable aléatoire $f(X)$ fonction de X. On commence par montrer que $(Y - u(X))$ est indépendante de X. Puisque le vecteur $[X, Y]'$ est gaussien et que :

$$\begin{bmatrix} X \\ Y - u(X) \end{bmatrix} = \begin{bmatrix} X \\ Y - (aX + b) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -a & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} 0 \\ -b \end{bmatrix} = A \begin{bmatrix} X \\ Y \end{bmatrix} + B$$

Le vecteur $(X, Y - u(X))'$ est gaussien aussi comme tranformée affine d'un vecteur gaussien, donc montrer l'indépendance de ses composantes revient à montrer leur décorrélation. Or

$$Cov(X, Y - u(X)) = Cov(X, Y) - Cov(X, u(X)) = Cov(X, Y) - Cov(X, aX + b) = 0,$$

donc X et $(Y - u(X))$ sont indépendantes, et reste aussi indépendante de toute fonction $f(X)$ de la variable X. Par suite

$$< f(X), Y - u(X) >= \mathbb{E}[f(X)(Y - u(X))] = \mathbb{E}[f(X)]\mathbb{E}[Y - u(X)] = 0.$$

Rappel . Le projeté orthogonal de Y sur un sous-espace vectoriel $F = Vect(e_1, ..., e_n)$, avec les e_i orthogonaux, est

$$\pi_F(Y) = \sum_{i=1}^{n} < Y, \frac{e_i}{\|e_i\|} > \frac{e_i}{\|e_i\|} = \sum_{i=1}^{n} \frac{< Y, e_i >}{\|e_i\|^2} e_i$$

Interprétation géométrique : Soit F le sous espace de $\mathbf{L}^2(X)$ engendré par le vecteur $(1, X)$.

Soit $(1, X - \mathbb{E}(X))$ base orthogonale de F :

$$< 1, X - \mathbb{E}(X) >= \mathbb{E}(1(X - \mathbb{E}(X))) = 0$$

Le premier vecteur donne $\dfrac{< Y, 1 >}{\mathbb{E}[1^2]} = \mathbb{E}[Y]$

le second vecteur donne $\dfrac{< Y, X - \mathbb{E}(X) >}{\mathbb{E}(X - \mathbb{E}(X))^2} = \dfrac{Cov(X, Y)}{Var(X)}$

On en déduit

$$\pi_F(Y) = \mathbb{E}[Y] + \frac{Cov(X, Y)}{Var(X)}(X - \mathbb{E}[X]) = \mathbb{E}[Y|X]$$

On retrouve bien la droite de régression ceci veut dire que dans le cas gaussien, la projection orthogonale de Y sur $\mathbf{L}^2(X)$ est exactement la projection orthogonale sur F=$vect(1, X)$.

1.4 Hyperplan de régression

Dans ce paragraphe, on ne fait aucune hypothèse de gaussianité. On suppose observer un V.a $X = (X_1, ..., X_n)$ de matrice de covariance Γ_X supposée de plein rang. On veut connaître la fonction affine des X_i, donc de la forme $f(X_1, ..., X_n) = b + a_1X_1 + ... + a_nX_n$ qui approche le mieux la variable aléatoire Y au sens des moindres carrés, autrement dit, l'erreur quadratique moyenne

$$\mathbb{E}[(Y - (b + a_1X_1 + ... + a_nX_n))^2]$$

soit minimale. Dans ce cas, on cherche l'hyperplan de régression, ceci revient à déterminer la projection $\pi_F(Y)$ de Y sur le sous-espace

$$F = Vect(1, X_1, ..., X_n).$$

Théoreme. *La projection orthogonale de la v.a Y sur F est*

$$\pi_F(Y) = b + \sum_{i=1}^{n} a_i(X_i - \mathbb{E}[X_i]) = \mathbb{E}[Y] + \Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X])$$

avec

$$\Gamma_{Y,X} = \mathbb{E}[(Y - \mathbb{E}[Y])(X - \mathbb{E}[X])'] = [Cov(Y, X_1), ..., Cov(Y, X_n)]$$

matrice ligne de covariance de la variable aléatoire Y et du vecteur aléatoire X.

preuve . la projection orthogonale de Y sur F est de la forme

$$\pi_F(Y) = b + \sum_{i=1}^{n} a_i X_i$$

par suite $Y - \pi_F(Y)$ est orthogonale à F est équivalent à dire que $Y - \pi_F(Y)$ est orthogonale à chacun des vecteurs qui engendrent F, c'est-à-dire :$1, X_1, ..., X_n$

L'orthogonalité à 1 donne

$$< Y - b - \sum_{i=1}^{n} a_i X_i, 1 >= \mathbb{E}[Y] - b - \sum_{i=1}^{n} a_i \mathbb{E}[X_i] = 0$$

c'est-à-dire

$$b = \mathbb{E}[Y] - \sum_{i=1}^{n} a_i \mathbb{E}[X_i]$$

L'orthogonalité à X_j donne les n équations

$$< Y - b - \sum_{i=1}^{n} a_i X_i, X_j >= 0 \quad 1 \leq j \leq n$$

ou encore

$$< Y - \mathbb{E}[Y] - \sum_{i=1}^{n} a_i(X_i - \mathbb{E}[X_i]), X_j >= 0 \quad 1 \leq j \leq n$$

Avec les notations de l'énoncé, ces n équations se résument sous forme matricielle à

$$\Gamma_{Y,X} = [a_1, ..., a_n]\Gamma_X$$

c'est-à-dire :

$$[a_1, ..., a_n] = \Gamma_{Y,X}\Gamma_X^{-1}$$

En revenant à $\pi_F(Y)$, ceci donne

$$\pi_F(Y) = b + \sum_{i=1}^{n} a_i X_i = \mathbb{E}[Y] - \Gamma_{Y,X}\Gamma_X^{-1}\mathbb{E}[X] + \Gamma_{Y,X}\Gamma_X^{-1}X$$

c'est-à-dire

$$\pi_F(Y) = \mathbb{E}[Y] + \Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X])$$

Remarque 2. *En prenant $X = X_1$, on retrouve bien la droite de régression puisque $\Gamma_{Y,X} = Cov(X,Y)$ et $\Gamma_X = Var(X)$.*

Corollaire 1. ***Erreur quadratique moyenne .*** *L'erreur quadratique moyenne dans l'approximation par l'hyperplan de régression, encore appelée variance résiduelle ou résidu, est*

$$\mathbb{E}[(Y - \pi_F(Y))^2] = \Gamma_Y - \Gamma_{Y,X}\Gamma_X^{-1}\Gamma_{X,Y}$$

avec $\Gamma_Y = Var(Y)$ et $\Gamma_{X,Y} = (\Gamma_{Y,X})'$.

preuve . Il suffit de l'écrire :

$$\mathbb{E}[(Y - \pi_F(Y))^2] = \mathbb{E}[((Y - \mathbb{E}[Y]) - \Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X]))^2]$$

ce qui donne une combinaison de 3 termes. Le premier est simple

$$\mathbb{E}[(Y - \mathbb{E}[Y])^2] = Var(Y).$$

Le deuxième l'est un peu moins

$$\mathbb{E}[(Y - \mathbb{E}[Y])\Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X])] = \Gamma_{Y,X}\Gamma_X^{-1}\mathbb{E}[(Y - \mathbb{E}[Y])(X - \mathbb{E}[X])] = \Gamma_{Y,X}\Gamma_X^{-1}\Gamma_{X,Y}$$

et le troisième encore moins

$$\mathbb{E}[(\Gamma_{Y,X}\Gamma_X^{-1}(X-\mathbb{E}[X]))^2] = \mathbb{E}[(\Gamma_{Y,X}\Gamma_X^{-1}(X-\mathbb{E}[X]))(\Gamma_{Y,X}\Gamma_X^{-1}(X-\mathbb{E}[X]))']$$

ce qui aboutit à

$$\mathbb{E}[(\Gamma_{Y,X}\Gamma_X^{-1}(X-\mathbb{E}[X]))^2] = \Gamma_{Y,X}\Gamma_X^{-1}\Gamma_{X,Y}.$$

On remet tout bout à bout

$$\mathbb{E}[(Y-\pi_F(Y))^2] = Var(Y) - 2\Gamma_{Y,X}\Gamma_X^{-1}\Gamma_{X,Y} + \Gamma_{Y,X}\Gamma_X^{-1}\Gamma_{X,Y} = Var(Y) - \Gamma_{Y,X}\Gamma_X^{-1}\Gamma_{X,Y}$$

1.5 Espérance conditionnelle gaussienne

On suppose maintenant le vecteur $(X_1, ..., X_n, Y)$ gaussien. L'éspérance conditionnel de Y sachant $X=[X_1, ..., X_n]'$ est la projection orthogonale de Y sur l'espace des fonctions $u(X)=u(X_1, ..., X_n)$, avec $u : \mathbb{R}^n \to \mathbb{R}$ telle que $\mathbb{E}[u^2(X)] < +\infty$. C'est la fonction qui minimise $\mathbb{E}[(Y-u(X))^2]$. On a vu que pour un vecteur gaussien bidimensionnel $[X, Y]'$, la droite de régression coïncide avec la courbe de régression. Plus généralement, on montre que pour un vecteur gaussien $[X_1, ..., X_n, Y]$, l'espérance conditionnelle coïncide avec la projection sur l'hyperplan de régression.

Théoreme. ***Espérance conditionnelle*** $\Leftrightarrow$ ***Hyperplan de régression .***

Si $[X_1, ..., X_n, Y]'$ *est un vecteur gaussien , alors*

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X_1, ..., X_n] = \mathbb{E}[Y] + \Gamma_{Y,X}\Gamma_X^{-1}(X-\mathbb{E}[X])$$

et la variance résiduelle vaut

$$\sigma^2 = \mathbb{E}[(Y-\mathbb{E}[Y|X])^2] = \Gamma_Y - \Gamma_{Y,X}\Gamma_X^{-1}\Gamma_{X,Y}$$

preuve . Notons $\pi_F(Y)$ la projection orthogonal de Y sur $F = Vect(1, X_1, ..., X_n)$ c'est-à-dire

$$\pi_F(Y) = \mathbb{E}[Y] + \Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X])$$

On a bien sûr $\pi_F(Y)$ de la forme $u(X)$ ci-dessus. De plus , dire que $(Y - \pi_F)$ est orthogonale au sous-espace F signifie que $(Y - \pi_F(Y))$ est décorrélée des variables X_i (puisque $(Y - \pi_F(Y))$ est centrée. Mais puisque tout est gaussien, c'est exactement dire que $(Y - \pi_F(Y))$ est indépendante du vecteur X. Pour toute fonction u, on a donc

$$\mathbb{E}[(Y - u(X))^2] = \mathbb{E}[((Y - \pi_F(Y))) + (\pi_F(Y) - u(X)))^2]$$

Ce qui donne

$$\mathbb{E}[(Y - u(X))^2] = \mathbb{E}[(Y - \pi_F(Y))^2] + 2\mathbb{E}[(Y - \pi_F(Y))(\pi_F(Y) - u(X))] + \mathbb{E}[(\pi_F(Y) - u(X))^2]$$

Or on vient de voir que

$$\mathbb{E}[(Y - \pi_F(Y))(\pi_F(Y) - u(X))] = 0$$

le troisième terme étant positif, donc pour toute fonction u, on a

$$\mathbb{E}[(Y - u(X))^2] \geq \mathbb{E}[(Y - \pi_F(Y))^2]$$

Remarque 3. *Le terme $\Gamma_{Y,X}\Gamma_X^{-1}\Gamma_{X,Y}$ correspond à la variance de la variable aléatoire $\mathbb{E}[Y|X]$: il est donc positif et par suite $\mathbb{E}[(Y - \mathbb{E}[Y|X])^2] \leq \Gamma_Y$. On a obtenu la décomposition orthogonale*

$$Y = \mathbb{E}[Y|X] + W = (\mathbb{E}[Y] + \Gamma_{Y,X}\Gamma_X^{-1}(X - \mathbb{E}[X])) + W$$

c'est-à-dire que $W = Y - \mathbb{E}[Y|X]$ est une variable aléatoire gaussienne indépendante des X_i. W est centré puisque $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$ et, par le théorème de Pythagore, sa variance est la variance résiduelle

$$\sigma^2 = \Gamma_Y - \Gamma_{Y,X}\Gamma_X^{-1}\Gamma_{X,Y}$$

En bref, on a

$$\begin{cases} W \sim \mathcal{N}(0, \sigma^2), \\ W \quad \perp X, \end{cases}$$

Théoreme. ***Cochran*** [1] **.** *Soit $X = (X_1, ..., X_n)'$ V.g centré réduit. Pour F un sous espace vectoriel de $\mathbb{R}^n$ de dimension p, on note P_F (resp.$P_{F^\perp}$) la projection orthogonale sur F (resp. $F^\perp$). Alors les vecteurs aléatoires $P_F X$ et $P_{F^\perp} X$ sont gaussiens indépendants de lois $P_F X \sim \mathcal{N}(0, P_F)$ et $P_{F^\perp} X \sim \mathcal{N}(0, P_{F^\perp})$, de plus, les variables aléatoires $\|P_F X\|^2$ et $\|P_{F^\perp} X\|^2$ sont indépendantes de lois $\|P_F X\|^2 \sim \chi^2_p$ et $\|P_{F^\perp} X\|^2 \sim \chi^2_{n-p}$.*

peuve . Le résultat est immédiat si on l'écrit dans une base orthonormée adaptée à la somme directe orthogonale $\mathbb{R}^n = F \oplus F^\perp$: soit $(u_1, ..., u_p)$ (resp.$(u_{p+1}, ..., u_n)$) une base orthonormée de F (resp.$F^\perp$), alors $u = (u_1, ..., u_n)$ est une base orthonormée de $\mathbb{R}^n$. Notons U la matrice (orthogonale, $U^t = U^{-1}$) de passage de la base canonique à la base u.

Les projections orthogonales sur F et $F^\perp$ s'expriment trés simplement dans la base u :$P_F = U I_p U^t$ et $P_{F^\perp} = U J_{n-p} U^t$, où I_p est la matrice diagonale avec des 1 sur les p premiers coefficients diagonaux et des 0 ensuite, et $J_{n-p} = I_d - I_p$.

On pose $Y = U^t X$. C'est encore un vecteur gaussien centré réduit (car il est de matrice de covariance $U^t I_d U = I_d$), qui correspond aux coordonnées de X dans la base u.

Pour Y, on a immédiatement que $I_p Y = (Y_1, ..., Y_p, 0, ..., 0)^t$ et $J_{n-p} Y = (0, ..., 0, Y_{p+1}, ..., Y_n)$ sont indépendants, de lois $\mathcal{N}(0, I_p)$ et $\mathcal{N}(0, J_{n-p})$, puis que $\|I_p Y\|^2 = \sum_{i=1}^p Y_i^2 \sim \chi^2_p$ et $\|J_{n-p} Y\|^2 = \sum_{i=p+1}^n Y_i^2 \sim \chi^2_{n-p}$.

On peut alors revenir au vecteur X en remarquant que $P_F X = U I_p Y$ et $P_{F^\perp} X = U J_{n-p} Y$ sont gaussiens centrés indépendants de matrice de covariance respective $U I_p U^t = P_F$ et $U J_{n-p} U^t = P_{F^\perp}$, puis, comme une transformation orthogonale préserve la norme, que

$$\|P_F X\|^2 = \|I_p Y\|^2 \sim \chi^2_p \quad \text{et} \quad \|P_{F^\perp} X\|^2 = \|J_{n-p} Y\|^2 \sim \chi^2_{n-p}$$

Chapitre 2

ANOVA

Un exemple de reproductibilité pour étudier les performances de trois laboratoires relativement à la détermination de la quantité de sodium de lasalocide dans de la nourriture pour de la volaille.

Une portion de norriture contenant la dose nominale de 85 mg/kg de sodium de lasalocide a été envoyée à chacun des laboratoires à qui il a été demandé de procéder à 10 réplications de l'analyse.

Les mesures de sodium de lasalocide obtenues sont exprimées en mg/kg. Elles sont reproduites dans le tableau suivant :

	Lab.B	**Lab.**C	**Lab.**D
1	87	88	85
2	88	93	84
3	84	88	79
4	84	89	86
5	87	85	81
6	81	87	86
7	86	86	88
8	84	89	83
9	88	88	83
10	86	93	83

Cette écriture de tableau est dite désempilée, nous pouvons l'écrire sous forme stan-

dard (empilée), c'est-à-dire avec deux colonnes, une pour le laboratoire et une pour la valeur de la teneur en sodium de lasalocide mesurée, et trente lignes pour chacune des observations réalisées.

Essai	**Laboratoire**	**Lasalocide**
1	Laboratoire B	87
2	Laboratoire B	88
3	Laboratoire B	84
4	Laboratoire B	84
5	Laboratoire B	87
6	Laboratoire B	81
7	Laboratoire B	86
8	Laboratoire B	84
9	Laboratoire B	88
10	Laboratoire B	86

Essai	**Laboratoire**	**Lasalocide**
1	Laboratoire C	88
2	Laboratoire C	93
3	Laboratoire C	88
4	Laboratoire C	89
5	Laboratoire C	85
6	Laboratoire C	87
7	Laboratoire C	86
8	Laboratoire C	89
9	Laboratoire C	88
10	Laboratoire C	93

Essai	**Laboratoire**	**Lasalocide**
1	Laboratoire D	85
2	Laboratoire D	84
3	Laboratoire D	79
4	Laboratoire D	86
5	Laboratoire D	81
6	Laboratoire D	86
7	Laboratoire D	88
8	Laboratoire D	83
9	Laboratoire D	83
10	Laboratoire D	83

Sur chaque **essai**, on observe **deux variables**.

1. Le laboratoire. Il est totalement contrôlé. La variable "Laboratoire" est considérée comme qualitative avec trois modalités bien déterminées : B, C, et D. Nous l'appelons le facteur. Ici, le facteur "Laboratoire" est à **effets fixes**.
2. La quantité de sodium de lasalocide. La variable "Lasalocide" est considérée comme quantitative comme généralement tous les résultats obtenus par une mesure. Nous l'appelons **la variable réponse**.

La variable mesurée dans un tel schéma expérimental sera notée Y. Pour les observations, nous utilisons deux indices :

1. Le premier indice , nous utiliserons en général l'indice i.
2. Le second indice, nous utiliserons en général l'indice j.

Ainsi, les observations seront notées en général :

$$y_{ij}, \quad i = 1, ..., I; \quad j = 1, ..., J$$

2.0.1 Définition et Notations

En se plaçant dans le cas équilibré, nous notons les moyennes de chaque échantillon par :

$$\overline{y}_{i,\bullet} = \frac{1}{J} \sum_{j=1}^{J} y_{ij}, \quad i = 1, ..., I$$

et les variances de chaque échantillon par :

$$s_i^2(y) = \frac{1}{J} \sum_{j=1}^{J} (y_{ij} - \overline{y}_{i,\bullet})^2, \quad i = 1, ..., I$$

Remarque 4. *Cette dernière formule exprime la variance non corrigée. Très souvent, dans les ouvrages ou logiciels, c'est la variance corrigée qui est utilisée : au lieu d'être divisée par J, la somme est divisée par $J - 1$.*

Après calculs avec le logiciel R, nous avons pour l'exemple cité plus haut :

$$\overline{y}_{1\bullet} = 85.5, \quad \overline{y}_{2\bullet} = 88.6, \quad \overline{y}_{3\bullet} = 83.8$$

$$s_{1,c} = 2.224 \quad s_{2,c} = 2.633, \quad s_{3,c} = 2.616.$$

Le nombre total d'observations est égale à : $n = I \times J = 3 \times 10 = 30$.

2.1 Objectif.

L'analyse de la variance **ANOVA** est une méthode statistique qui permet d'étudier la modification de la moyenne μ d'une quantité Y (variable réponse quantitative) selon l'influence éventuelle d'un ou de plusieurs facteurs d'expérience qualitatifs (traitements ...). Dans le cas où la moyenne n'est influencée que par un seul facteur (noté facteur A), il s'agit d'une analyse de la variance à un seul facteur ("one way ANOVA"). Un facteur A est souvent une variable qualitative présentant un nombre restreint de modalités. Le nombre de modalités du facteur A sera noté I. Nous supposons que Y suit une loi normale $\mathcal{N}(\mu_i, \sigma^2)$ sur chaque sous-population i définie par les modalités de A. L'objectif est ici de tester l'égalité des moyennes de ces populations, à savoir de tester l'hypothèse nulle :

$$H_0 : \mu_1 = \mu_2 = ... = \mu_I$$

contre l'hypothèse alternative :

$H_1 : \exists i_0 \neq i^{'}$ tel que $\mu_{i_0} \neq \mu_{i^{'}}$ (il existe au moins deux moyennes différentes).

Nous regroupons les valeurs que peut prendre la réponse Y dans les facteurs A_i lors des J répétitions dans le tableau suivant :

Facteur A	Y
A_1	Y_{11},...,Y_{1J}
.	
A_i	Y_{i1},...,Y_{iJ}
.	
A_I	Y_{I1},...,Y_{IJ}

2.1.1 Notations .

Nous avons observé $n = I * J$ valeurs de la variable Y indexée par deux indices i et j. La moyenne de ces valeurs par rapport à l'indice i est notée $Y_{\bullet j}$. Il s'agit simplement de

la moyenne de valeurs de la j-ème colonne du tableau :

$$Y_{\bullet j} = \frac{1}{I}\sum_{i=1}^{I} Y_{ij}$$

La moyenne de ces valeurs par rapport à l'indice j est notée $Y_{i\bullet}$. Il s'agit simplement de la moyenne de valeurs de la i-ème ligne du tableau :

$$Y_{i\bullet} = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$$

La moyenne globale par rapport aux indices i et j est notée $Y_{\bullet\bullet}$. Il s'agit simplement de la moyenne globale du tableau :

$$Y_{\bullet\bullet} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij},$$

Là aussi, il est aisé d'avoir les lois des éléments ci-dessus. Par exemple : $Y_{i\bullet} \sim \mathcal{N}(\mu_i, \sigma^2/J)$ et $Y_{\bullet\bullet} \sim \mathcal{N}(\mu, \sigma^2/n)$ où $n = I * J$.

2.2 Equation d'analyse de variance et tests .

Alors, la variation totale théorique, ou somme totale des carrés des écarts est égale à :

$$SC_{TOT} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{\bullet\bullet})^2 \tag{2.1}$$

Nous appellons variation théorique due au facteur A, la quantité :

$$SC_F = J\sum_{i=1}^{I}(Y_{i\bullet} - Y_{\bullet\bullet})^2 \tag{2.2}$$

et variation résiduelle, la quantité :

$$SC_R = \sum_{i=1}^{I}\left(\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet})^2\right) \tag{2.3}$$

d'où la formule

$$\boxed{SC_{TOT} = SC_F + SC_R}$$

preuve . Remarquons que : $Y_{ij} - Y_{\bullet\bullet} = Y_{ij} - Y_{i\bullet} + Y_{i\bullet} - Y_{\bullet\bullet}$. D'où :

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{\bullet\bullet})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i\bullet} - Y_{\bullet\bullet})^2$$

car le double produit, dans le développement de la somme au carré, vaut zéro. En effet :

$$2\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet})(Y_{i\bullet} - Y_{\bullet\bullet}) = 2\sum_{i=1}^{I}(Y_{i\bullet} - Y_{\bullet\bullet})\left[\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet})\right]$$

D'autre part, nous remarquons que : $\sum_{j=1}^{J} Y_{ij} = JY_{i\bullet}$ par définition de $Y_{i\bullet}$. D'où :

$$\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet}) = JY_{i\bullet} - JY_{i\bullet} = 0$$

D'où le résultat.

Proposition 10. *Sous les hypothèses de normalité et d'égalité des variances, on a :*

$$\frac{SC_F}{\sigma^2} \sim \chi^2_{I-1}$$

Sous les hypothèses usuelles de l'analyse de la variance, nous avons :

$$\frac{SC_R}{\sigma^2} \sim \chi^2_{n-I}$$

Sous les hypothèses habituelles de l'analyse de la variance, les statistiques SC_F et SC_R sont indépendantes et on a :

$$\frac{SC_T}{\sigma^2} \sim \chi^2_{n-1}$$

Preuve .[2]

1. *Il est facile de voir que $SC_F = \sum_{i=1}^{I} JY_{i\bullet} - nY_{\bullet\bullet}^2$. Posons $Z_i = \sqrt{J}Y_{i\bullet}$ pour $i = 1, ..., I$.*

   Nous avons alors : $Z_i \sim \mathcal{N}(\sqrt{J}\mu_i, \sigma^2)$ et $Y_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{I}\sqrt{J}Z_i$.

   Nous en déduisons que :

$$\begin{aligned} SC_F &= \sum_{i=1}^{I} Z_i^2 - (\frac{1}{\sqrt{n}}\sum_{i=1}^{I}\sqrt{J}Z_i)^2 \\ &= \sum_{i=1}^{I}(1 - \frac{J}{n})Z_i^2 - 2\sum_{i<j}\frac{\sqrt{J}\sqrt{J}}{n}Z_iZ_j \end{aligned}$$

SC_F s'exprime alors comme une forme quadratique en $\boldsymbol{Z} = (Z_1, ..., Z_I)'$.

La matrice A associée à cette forme quadratique s'exprime sous la forme $A = I_d - \frac{1}{n}\nu\nu'$ où $\nu = (\sqrt{J}, \sqrt{J}, ..., \sqrt{J})'$.

Calculons A^2. On a ici :

$$\begin{aligned} A^2 &= (I_d - \frac{1}{n}\nu\nu') \times (I_d - \frac{1}{n}\nu\nu') \\ &= I_d - \frac{1}{n}\nu\nu' - \frac{1}{n}\nu\nu' + \frac{1}{n^2}\nu\nu'\nu\nu' \end{aligned}$$

Nous pouvons remarquer que $\nu\nu' = \sum_{i=1}^{I} \sqrt{J}\sqrt{J} = n$, donc les deux derniers termes de la partie droite de la dernière équation s'annulent et on a $A^2 = A$

D'où $\frac{SC_F}{\sigma^2}$ suit alors χ^2_d où d est l'ordre de la matrice A.

2. *En effet, posons :*

$$S_i^2 = \frac{1}{J-1}\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet})^2.$$

Pour $i = 1, ..., I$, on a $(J-1)\frac{S_i^2}{\sigma^2} \sim \chi^2_{J-1}$. Par indépendante des échantillons, $\frac{SC_R}{\sigma^2} = \sum_{i=1}^{I}(J-1)\frac{S_i^2}{\sigma^2} \sim \chi^2_{n-I}$ puisque $\sum_{i=1}^{I}(J-1) = n - I$.

3. *Posons $u = (1, 1, ..., 1)' \in \mathbb{R}^n$, $Z = (Y_{ij})_{i=1,...,I, j=1,...,J}$*

Soit l'espace $F = vect(u)$ tel que $\dim F = 1$, d'aprés le théoreme de Cochran nous avons $P_F(Y_{ij})$ et $P_{F^\perp}(Y_{ij})$ sont indépendants .

En effet : $P_F(Z) = \frac{(Z,u)u}{\|u\|^2} = \left(\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I} Y_{ij}\right) u = \overline{Y}u$ et $P_{F^\perp}(Z) = Z - \overline{Y}u$

D'où $\frac{1}{\sigma^2}\|P_{F^\perp}(Z)\|^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(Y_{ij} - \overline{Y})^2}{\sigma^2} = \frac{SC_T}{\sigma^2} \sim \chi^2_{n-1}$

Table de l'analyse de variance .

$$\frac{1}{I \times J}\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{\bullet\bullet})^2 = \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - Y_{i\bullet})^2 + \frac{1}{I \times J}\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i\bullet} - Y_{\bullet\bullet})^2 \qquad (2.4)$$

que nous pouvons noter :

$$S^2 = S_F^2 + S_R^2$$

appelée formule de l'analyse de la variance, où S_F^2 représente la variance due au facteur A, où S_R^2 représente la variance résiduelle, et S^2 la variance globale.

Les diverses quantités (4.2), (4.3) et (4.4) suivent, au facteur multiplicatif σ^2 prés, des lois du khi-deux avec des nombres de degrés de liberté respectifs $IJ-1$, $I-1$ et $IJ-I$. Ceci est résumé dans le tableau suivant :

Source	**Variations**	**Degrés de liberté**	**Carrées moyens**	F
Facteur	SC_F	$I-1$	$S_F^2=\frac{SC_F}{I-1}$	$F=\frac{S_F^2}{S_R^2}$
Résidu	SC_R	$IJ-I=I(J-1)$	$S_R^2=\frac{SC_R}{I(J-1)}$	
Total	SC_{TOT}	$I(J-1)$	$S_T^2=\frac{SC_{TOT}}{I(J-1)}$	

Exemple .

Nous voulons comparer des hauteurs moyennes, exprimées en mètres, des arbres de trois types de hêtraies. Nous cherchons effectivement à savoir s'il existe ou non, en moyenne, des différences significatives de hauteurs d'arbres entre les trois types de forêts. Nous supposons que les hypothèses de normalitè et d'ègalité des variances sont satisfaites. Les données sont fournies dans le tableau suivant :

Type 1	Type 2	Type 3
23.4	22.5	18.9
24.4	22.9	21.1
24.6	23.7	21.2
24.9	24.0	22.1
25.0	24.4	22.5
26.2	24.5	23.6
26.3	25.3	24.5
26.8	26.0	24.6
26.8	26.2	26.2
26.9	26.4	26.7
27.0	26.7	
27.6	26.9	
27.7	27.4	
	28.5	

Nous pouvons aussi réaliser cette étude sous R. Les commandes sont les suivantes :

>hetraie<-rep$(1:3, c(13, 14, 10))$

>hauteur<-$c(23.4, 24.4, 24.6, 24.9, 25.0, 26.2, 26.3, 26.8, 26.8, 26.9, 27.0, 27.6, 27.7, 22.5, 22.9, 23.7, 24.0,$

+24.4, 24.5, 25.3, 26.0, 26.2, 26.4, 26.7, 26.9, 27.4, 28.5, 18.9, 21.1, 21.2, 22.1, 22.5, 23.6, 24.5, 24.6, 26.2, 26.7)

>hetraie<-factor(hetraie)

>arbre<-data.frame(hetraie,hauteur)

>modele1<-aov(hauteur hetraie,data=arbre)

> summary(modele1)

Le tableau d'analyse de variance fourni est le suivant :

	Df	Sum Sq	Mean Sq	F value	Pr($> F$)
hetraie	2	48.88	24.441	7.124	0.00261
Residuals	34	116.65	3.431		

A un seuil égale à 5% l'hypothèse d'égalité des hauteurs moyennes des arbres dans les trois types de hêtraies est rejetée.

2.3 Analyse de la variance à deux facteurs .

L'analyse de variance à deux facteurs peut être considérée comme une généralisation de l'analyse de variance à un facteur, permettant de tenir compte simultanément de deux facteurs. Les deux facteurs peuvent être placés soit sur un pied d'égalité, soit subordonnés l'un à l'autre. Dans le premier cas, les modèles d'analyse de variance sont dits croisés, et, dans le second cas, ils sont appelés hiérarchisés ou multi-niveaux.

2.4 Equation d'analyse de variance et tests .

Nous regroupons les valeurs prises par la variable réponse Y dans les conditions (A_i, B_j) dans le tableau ci-dessous :

	Facteur B				
Facteur A	B_1	...	B_j	...	B_J
A_1	Y_{11}	...	Y_{1j}	...	Y_{1J}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_i	Y_{i1}	...	Y_{ij}	...	Y_{iJ}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_I	Y_{I1}	...	Y_{Ij}	...	Y_{IJ}

La variation théorique totale est définie par :

$$SC_{TOT} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - Y_{\bullet,\bullet})^2 \tag{2.5}$$

Nous appelons variation théorique due au facteur A, la quantité :

$$SC_A = J\sum_{i=1}^{I}(Y_{i,\bullet} - Y_{\bullet,\bullet})^2 \tag{2.6}$$

De la même façon, la variation théorique due au facteur B est :

$$SC_B = I\sum_{j=1}^{J}(Y_{\bullet,j} - Y_{\bullet,\bullet})^2 \tag{2.7}$$

La variance résiduelle est définie par :

$$SC_R = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - Y_{i,\bullet} - Y_{\bullet,j} + Y_{\bullet,\bullet})^2 \tag{2.8}$$

Nous montrons alors aisément la relation fondamentale de l'analyse de variance à deux facteurs sans répétition :

$$SC_{TOT} = SC_A + SC_B + SC_R \tag{2.9}$$

Source	**Degrés de liberté**
Facteur A	n_A=$I-1$
Facteur B	n_B=$J-1$
Résiduelle	n_R=$(I-1)(J-1)$
Totale	n_{TOT}=$IJ-1$

Aux différentes sommes des carrés des écarts peuvent être associés des nombres de degrés de liberté :

$S_A^2 = \frac{SC_A}{n_A}$; $S_B^2 = \frac{SC_B}{n_B}$; $S_R^2 = \frac{SC_R}{n_R}$; $S_T^2 = \frac{SC_{TOT}}{n_{TOT}}$

qui constituent eux aussi des mesures globales de variations.
Notons y des données expérimentales $y_{1,1}, ..., y_{1,J}, y_{2,1}, ..., y_{2,J}, ..., y_{I,J}$ permettant une réalisation du tableau précédent tableau

La variation totale observée sur la liste y de données expérimentales est définie par :

$$sc_{TOT} = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{i,j} - y_{\bullet,\bullet})^2 \tag{2.10}$$

La variation due au facteur A observée sur la liste y de données expérimentales est définie par :

$$sc_A = J\sum_{i=1}^{I}(y_{i,\bullet} - y_{\bullet,\bullet})^2 \tag{2.11}$$

La variation due au facteur B observée sur la liste y de données expérimentales est définie par :

$$sc_B = I\sum_{j=1}^{J}(y_{\bullet,j} - y_{\bullet,\bullet})^2 \tag{2.12}$$

La variation résiduelle observée sur la liste y de données expérimentales est quant à elle égale à:

$$sc_R = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{i,j} - y_{i,\bullet} - y_{\bullet,j} + y_{\bullet,\bullet})^2 \tag{2.13}$$

La relation fondamentale de l'analyse de variance reste valable lorsqu'elle est évaluée sur la liste y de données expérimentales :

$$sc_{TOT} = sc_A + sc_B + sc_R \tag{2.14}$$

Nous désirons faire les tests d'hypothèses suivantes:

$$H_0^{'} = \alpha_1 = \alpha_2 = ... = \alpha_I = 0$$

contre

$$H_1^{'} : \exists i_0 \in \{1, 2, ..., I\} \backslash \alpha_{i0} \neq 0$$

Nous pouvons répéter tout ce qui précède pour le facteur B. Nous pouvons souhaiter tester les hyphothèses:

$$H_0^{''} = \beta_1 = \beta_2 = ... = \beta_I = 0$$

contre

$$H_1^{''} : \exists j_0 \in \{1, 2, ..., J\} \backslash \beta_{j0} \neq 0$$

Le tableau d'analyse de la variance à deux facteurs résume les choses ci-dessous :

Source	**Variations**	**d.d.l**	**Carrés moyens**	F	**Décision**
Facteur A	sc_A	n_A	$s_A^2 = \frac{sc_A}{n_A}$	$f_A = \frac{s_A^2}{s_R^2}$	$H_0^{'}$ ou $H_1^{'}$
Facteur B	sc_B	n_B	$s_B^2 = \frac{sc_B}{n_B}$	$f_B = \frac{s_B^2}{s_R^2}$	$H_0^{''}$ ou $H_1^{''}$
Résiduelle	sc_R	n_R	$s_R^2 = \frac{sc_R}{n_R}$		
Totale	sc_{TOT}	n_{TOT}			

Exemple . L'influence d'un traitement grossissant, à base de vitamines, est étudiée sur des animaux de races différentes. Pour cela nous disposons

d'animaux de trois races, notées R_i , pour $i = 1, 2, 3$, et nous avons effectué trois traitements, notés D_j , pour $j = 1, 2, 3$, utilisant respectivement $5, 10$ et 15μg de vitamines $B12$ par $cm3$. Le gain moyen de poids par jour est mesuré, à l'issue d'un traitement de 50 jours dans chaque cas. Un seul animal est utilisé pour chaque couple $\ll$ race-traitement $\gg$.

Traitement\Race	R_1	R_2	R_3
D_1	1.26	1.21	1.19
D_2	1.29	1.23	1.23
D_3	1.38	1.27	1.22

L'objectif est d'effectuer une analyse de la variance à deux facteurs sans répétition (il y a en effet une seule observation par $\ll$ case$\gg$). Les facteurs, contrôlés, à effets fixes, sont la race et la dose, tous les deux à 3 modalités. La réponse est le gain moyen de poids.
Nous désirons tester les hypothèses suivantes:

H_0^R: Les races n'ont pas d'effets sur la prise de poids
H_1^R: Les races ont un effet sur la prise de poids

Puis

H_0^D: Les doses n'ont pas d'effet sur la prise de poids
H_1^D: Les doses ont un effet sur la prise de poids

Le tableau d'analyse de la variance correspondant est :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
races	2	0.015267	0.007633	9.745	0.0290
doses	2	0.007400	0.003700	4.723	0.0885
Resudual	4	0.003133	0.000783		

Nous décidons alors que :

1. H_1^R est vraie : il y a un effet de la race sur le gain de poids (p=$0,0290$).
2. H_0^D est vraie : il n' y a pas d'effet de la dose sur le gain de poids (p=$0,0885$).

Chapitre 3

Analyse en composantes principales

Introduction

L'ACP est une méthode de base en statistique exploratoire multidimensionnelle, elle permet de représenter graphiquement les corrélations entre variables et les ressemblances entre individus, elle transforme un grand nombre de variables corrélées en un plus petit nombre de variables indépendantes les unes des autres appelées les composantes principales.

3.1 Les données en ACP .

En ACP Les données sont les mesures effectuées sur n unités $\{u_1, u_2, ..., u_i, ...u_n\}$. Les p variables quantitatives qui représentent ces mesures sont $\{v_1, v_2, ..., v_j, ...v_p\}$, elle consiste à projeter les points sur une droite, un plan...un sous-espace à q dimensions (avec $q < p$) choisi de façon à optimiser un certain critère.

$$X = \begin{array}{c} \\ u_1 \\ u_2 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{array} \begin{array}{c} \begin{array}{cccccc} v_1 & v_2 & . & v_j & . & . & v_p \end{array} \\ \begin{pmatrix} x_{11} & . & . & . & . & . & x_{1p} \\ x_{21} & . & . & . & . & . & x_{2p} \\ \vdots & & & \vdots & & & \vdots \\ x_{i1} & . & x_{ij} & . & . & . & x_{ip} \\ \vdots & & & \vdots & & & \vdots \\ x_{n1} & . & . & . & . & . & x_{np} \end{pmatrix} \end{array}$$

On note :
L'individu u_i : élément de $\mathbb{R}^p$.

Variable v_j: élément de $\mathbb{R}^n$.
x_{ij}: est la valeur prise par la variable j sur l'individu i.

Avec $U_i^t = (x_{i1} \quad x_{i2} \quad . \quad x_{ij} \quad . \quad . \quad x_{ip})$; Ce qui donne $U_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ . \\ x_{ij} \\ . \\ . \\ x_{ip} \end{pmatrix}$

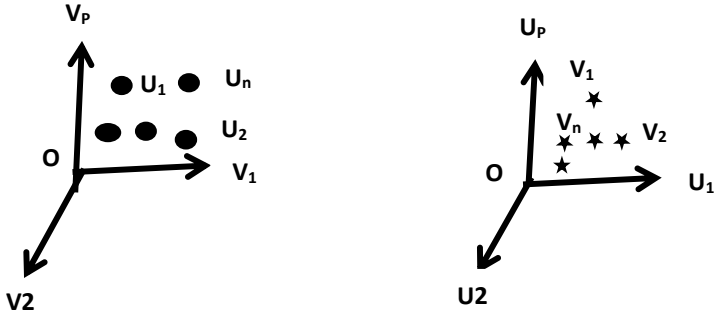
et $V_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ . \\ x_{ij} \\ . \\ . \\ x_{nj} \end{pmatrix}$

Nous constatons que ces espaces etant de dimension plus en génèral à 2 et mème 3, nous pouvons visualiser ces reprèsentations.
Dans ce cas, nous cherchons une représentation des n individus dans un sous espace F_q de $\mathbb{R}^p$ de dimension q et c'est **le principe d'une ACP.**
Autrement dit nous cherchons à définir q nouvelles variables combinaisons linéaires de p variables initiales et qui nous ferons perdre le moindre d'information possible. Nous avons

$$C_l = \sum_{i=1}^{p} a_{li} X_i$$

- C_l sont appelées **composantes principales.**
- axes qu'elles déterminent sont appelés **axes principaux.**
- formes linéaires associées sont appelées **facteurs principaux.**

Remarque . Les a_{li} représentent les éléments du vecteur propre associé à la matrice de variance-covariance.



3.2 Choix d'une distance .

Il faut choisir une distance pour faire une représentation géométrique donc la distance entre deux unité u_i et $u_i^{'}$ utilisée par l'ACP dans l'espace est égale à:

$$d^2(u_i, u_i^{'}) = \sum_{j=1}^{p}(x_{ij} - x_{i^{'}j})^2,$$

avec cette distance, toutes les variables jouent le même rôle et les axes définis par les variables constituent une base orthogonale. à cette distance on associe: $< \overrightarrow{ou_i}, \overrightarrow{ou_{i'}} >= \sum_{j=1}^{p} x_{ij}x_{i^{'}j} = U_i^t U_i$, ainsi $\|\overrightarrow{ou_i}\|^2 = \sum_{j=1}^{p} x_{ij}^2 = U_i^t U_i$.
Nous pouvons alors définir l'angle α entre deux vecteurs par son cosinus

$$\cos(\alpha) = \frac{< \overrightarrow{ou_i}, \overrightarrow{ou_{i'}} >}{\|\overrightarrow{ou_i}\|\|\overrightarrow{ou_{i'}}\|} = \frac{\sum_{j=1}^{p} x_{ij}x_{i^{'}j}}{\sqrt{\sum_{j=1}^{p} x_{ij}^2 \sum_{j=1}^{p} x_{i^{'}j}^2}} = \frac{U_i^t U_{i'}}{\sqrt{(U_i^t U_i)(U_{i'}^t U_{i'})}}$$

3.3 Choix de l'origine .

Il faut choisir une origine liée au nuage de point car le point o correspondant au vecteur de coordonnées toutes nulles n'est pas une origine satisfaisante, car si les coordonnées des points du nuage des individus sont grandes, le nuage est èloignè de cette origine.
Le centre de gravitè est dèfini par: $\sum_{i=1}^{n} p_i \overrightarrow{Gu_i} = 0$, avec $\sum_{i=1}^{n} p_i = 1$. Ces poids sont représentés par une matrice diagonale D taille n. Si $p_i = \frac{1}{n}$, nous avons

$$G = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{pmatrix} = \begin{pmatrix} \overline{x_{.1}} \\ \overline{x_{.j}} \\ \overline{x_{.p}} \end{pmatrix}$$

Prendre G comme origine, revient alors à travailler sur les tableau des données centrées:

$$X_c = \begin{pmatrix} x_{11} - \overline{x_{.1}} & . & x_{1j} - \overline{x_{.j}} & . & x_{1p} - \overline{x_{.p}} \\ . & . & . & . & . \\ x_{i1} - \overline{x_{.1}} & . & . & . & x_{ip} - \overline{x_{.p}} \\ . & . & . & . & . \\ x_{n1} - \overline{x_{.1}} & . & x_{nj} - \overline{x_{.j}} & . & x_{np} - \overline{x_{.p}} \end{pmatrix}$$

et le vecteur de coordonnées centrées de l'unité u_i est $U_{ci} = \begin{pmatrix} x_{i1} - \overline{x_{.1}} \\ . \\ x_{ij} - \overline{x_{.j}} \\ . \\ x_{ip} - \overline{x_{.p}} \end{pmatrix}$

celuis des coordonnées centrées de la variable v_j est $V_{cj} = \begin{pmatrix} x_{1j} - \overline{x_{.j}} \\ . \\ x_{ij} - \overline{x_{.j}} \\ . \\ x_{nj} - \overline{x_{.j}} \end{pmatrix}$

3.4 Inertie total .

Nous notons I_T le moment d'inertie du nuage des individus par rapport au centre de gravité

$$I_T = \frac{1}{n}\sum_{i=1}^{n} d^2(G, u_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \overline{x_{.j}})^2 = \frac{1}{n}\sum_{i=1}^{n} U_{ci}^t U_{ci}$$

Si ce moment d'inertie est grand cela signifie que le nuage est très dispersé sinon il est très concentré sur son centre de gravité.
Remarque . I_T peut aussi s'écrire sous la forme

$$I_T = \sum_{j=1}^{p}[\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_{.j}})^2] = \sum_{j=1}^{p} Var(v_j)$$

où $Var(v_j)$ est la variance empirique de la variable v_j.
En notant Σ la matrice de variance-covariance

$$\Sigma = \begin{pmatrix} S_1^2 & . & . & . & S_{1p} \\ . & . & . & . & . \\ . & . & S_j^2 & . & . \\ . & . & . & . & . \\ S_{p1} & . & . & . & S_p^2 \end{pmatrix}$$

avec S_k est l'écart type, Donc $I_T = Tr(\Sigma)$.
Dans le cas où les variables sont centrèes rèduites la variance de chaque variable vaut 1.

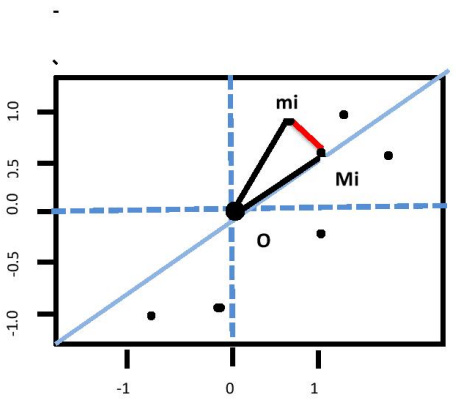
3.5 Décomposition de l'inertie totale .

Soit un vecteur unitaire u dans $\mathbb{R}^p$, il définira un axe principale que nous allons chercher. Le point M_i se projette sur cet axe en m_i. On a : $M_i = m_i + (M_i - m_i)$ et $\|M_i\|^2 = \|m_i\|^2 + \|M_i - m_i\|^2$

$$\frac{1}{n}\sum_{i=1}^{n}\|M_i\|^2 = \frac{1}{n}\sum_{i=1}^{n}\|m_i\|^2 + \frac{1}{n}\sum_{i=1}^{n}\|M_i - m_i\|^2$$

$$I_T = I_S(u) + I_M(u)$$

où $I_S(u)$ à maximiser et $I_M(u)$ à minimiser .



Recherche du vecteur directeur .[4] La matrice X_0 contient les n points centrés.

$$X_0 = \begin{pmatrix} x_1 - m(X) & y_1 - m(Y) \\ \vdots & \vdots \\ x_n - m(X) & y_n - m(Y) \end{pmatrix}$$

Le vecteur u recherché est unitaire. Nous l'écrivons sous la forme $u = \begin{pmatrix} a \\ b \end{pmatrix}$ avec $a^2 + b^2 = 1$
L'inertie statistique ou l'inertie projetée est :

$$I_S(u) = \frac{1}{n}\sum_{i=1}^{n} \|m_i\|^2 = \frac{1}{n}(X_0 u)^t X_0 u = u^t(\frac{1}{n}X_0^t X_0)u$$

$$= \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

$$= a^2\sigma_X^2 + 2ab\sigma_{XY} + b^2\sigma_Y^2$$

La matrice $\begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$ est dite matrice de variances-covariances des deux variables. Nous la notons $C = \frac{1}{n}X_0^t X_0$, elle est symétrique. Son polynôme caractéristique s'écrit

$$|C - \lambda I| = \begin{vmatrix} \sigma_X^2 - \lambda & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 - \lambda \end{vmatrix} = \lambda^2 - \lambda(\sigma_X + \sigma_Y) + \sigma_X\sigma_Y - \sigma_{XY}^2$$

- Le pôlynome caractéristique a toujours deux racines donc C a toujours deux valeurs propres et deux vecteurs propres .
- Les valeurs propres sont en général distinctes. On les note λ_1 et λ_2.

1. $\lambda_1 + \lambda_2 = \sigma_X + \sigma_Y$
2. $\lambda_1 \times \lambda_2 = \sigma_X \times \sigma_Y - \sigma_{XY}^2$

Toute matrice symétrique admet une base de vecteurs propres orthogonaux. Donc,

$$C = U \wedge U^t$$

avec: $U = \begin{pmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \end{pmatrix}$ et $\wedge = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$

$$\begin{aligned} I_S(u) &= u^t C u \\ &= \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \\ &= \begin{bmatrix} \alpha & \beta \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \\ &= \lambda_1\alpha^2 + \lambda_2\beta^2 \leq \lambda_1\alpha^2 + \lambda_1\beta^2 = \lambda_1 \end{aligned}$$

$\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ représente les coordonnées du vecteur u dans la base des vecteurs propres.

L'inertie ne peut dépasser la première valeur propre et l'atteint pour $\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

Généralisation à p variables . $C = U \wedge U^t$ admet une base de p vecteurs propres orthonormés.

- Le premier vecteur propre normé u_1 est un vecteur de $\mathbb{R}^p$ qui maximise l'inertie projetée.
- Le deuxième vecteur propre normé u_2 est un vecteur de $\mathbb{R}^p$,orthogonal à u_1, qui maximise à nouveau l'inertie projetèe.
- et ainsi de suite pour $u_3 ... u_q$ axes suivants où q représente le rang de la matrice diagonalisée.

Coordonnées des projections.[4]
Si u_k est le vecteur propre de rang k, les coordonnées des projections des n points sont obtenus simplement par :

$$l_k = \begin{bmatrix} l_{1k} \\ \vdots \\ l_{nk} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{p} (x_{1j} - \overline{x_{.j}}) u_{jk} \\ \vdots \\ \sum_{j=1}^{p} (x_{nj} - \overline{x_{.j}}) u_{jk} \end{bmatrix}$$

Soit en écriture matricielle: $l_k = X_0 u_k$
Axes principaux et coordonnées .

- u_k est appelé vecteur directeur déterminant l'axe principale F_k .
- l_k est appelée vecteur des coordonnées sur l'axe principale F_k. C'est une variable artificielle de moyenne nulle et de variance λ_k

$$\begin{aligned} m(I_k) &= \frac{1}{n}\sum_{i=1}^{n} l_{ik} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{1j} - \overline{x_{.j}})u_{jk} \\ &= \sum_{j=1}^{p} u_{jk}\sum_{i=1}^{n}(x_{1j} - \overline{x_{.j}}) = 0 \end{aligned}$$

$$\begin{aligned} v(I_k) &= \frac{1}{n}\sum_{i=1}^{n} l_{ik}^2 = \frac{1}{n}(X_0 u_k)^t X_0 u_k = u_k^t C u_k \\ &= \lambda_k u_k^t u_k = \lambda_k \end{aligned}$$

Remarque : La somme des valeurs propres est l'inertie totale .

3.6 Calcul des covariances et des corrélations .

Si nous voulons travailler avec des variables centrées et réduites, nous passons du tableau des valeurs centrées au tableau des valeurs centrées réduites de la façon suivante

$$X_{cv} = X_c (D_\Sigma)^{-\frac{1}{2}}.$$

avec $(D_\Sigma)^{-\frac{1}{2}} = \begin{pmatrix} \frac{1}{\sigma_1} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \frac{1}{\sigma_j} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \frac{1}{\sigma_p} \end{pmatrix}$

Si nous calculons la matrice de covariance à partir d'un tableau de données centrées et réduites on obtient la matrice de corrélations empiriques $R = \frac{1}{n} X_{cv}^t X_{cv} = (D_\Sigma)^{-\frac{1}{2}} \Sigma (D_\Sigma)^{-\frac{1}{2}}$

3.7 ACP par projection .

Nous donnons maintenant une généralisation de l'exemple cité plus haut. Nous nous basons sur la méthode de Lagrange pour déterminer les q vecteurs directeurs des axes principaux.
Soit la matrice $X_{n\times p}$, chaque ligne représente un individu et chaque colonne représente une variable. Nous supposerons que chaque variable est centrée. Nous définissons maintenant u_1 le vecteur unitaire (i.e: de norme 1 et $u_1 u_1^{'} = 1$).
C'est le vecteur présentant la plus grande dispersion des projections. Les projections des n observations sur le vecteur u_1 sont données par $P_{u_1}(X) = Xu_1$. (Produit scalaire)
La somme des carrés de ces projections (inertie) est:

$$P_{u_1}^{'}(X) P_{u_1}(X) = u_1^{'} X^{'} X u_1.$$

Pour le premier vecteur propre, on cherche un vecteur unitaire u^* tel que

$$u^* = \arg \max_{\{u_1 \in \mathbb{R}^n, u_1 u_1^{'}=1\}} u_1^{'} X^{'} X u_1$$

Nous cherchons donc le vecteur u^* tel que la projection du nuage sur u^* ait une inertie (ou une variance) maximale. En introduisant les multiplicateurs de Lagrange pour s'affranchir de la contrainte dans le problème de maximisation.

$$(u^*, \lambda) = \arg \max_{\{u_1 \in \mathbb{R}^n, u_1 u_1^{'}=1\}} u_1^{'} X^{'} X u_1 - \lambda(u_1^{'} u_1 - 1)$$

La solution est la racine de la dérivée de l'expression ci-dessous

$$\begin{cases} 2[X^{'} X u_1 - \lambda u_1] = 0 \\ u_1 u_1^{'} = 1 \end{cases}$$

Simplifiant, nous trouvons

$$\begin{cases} X^{'} X u_1 = \lambda u_1 \\ u_1 u_1^{'} = 1 \end{cases}$$

Si nous remarquons maintenant que

$$\max(u_1^{'} X^{'} X u_1) = \max(u_1^{'} \lambda u_1) = \max \lambda = \lambda_1$$

Nous avons que le premier axe factoriel u^* est associé à la plus grande valeur propre de $X^{'}X$.

Remarque

1. Nous savons que la somme des valeurs propres d'une matrice est égale à la trace de la matrice originale. De plus $\lambda_1 / \sum_{i=1}^{n} \lambda_i$ indique la proportion de la variance totale prise en charge par le 1^{er} vecteur propre.

3.8 Analyse dans l'espace des échantillons .

Nous nous sommes intéressés, jusqu'à maintenant, au nuage des n observations dans l'espace des p variables. Nous pourrions également considérer le nuage de p variables dans l'espace des n observations. Nous cherchons le sous-espace de dimension s ($s \leq p$) pour lequel la somme des carrés des projections est maximale.

Nous appliquons la même technique que précédemment et nous trouvons que la solution, pour le premier vecteur est donnée par le système suivant:

$$X X^{'} v_1 = \beta_1 v_1$$

$$v_1^{'} v_1 = 1$$

Le premier vecteur propre de $XX^{'}$ est celui qui maximise la variance des projections. Comme tantôt, le plan maximisant la variance des projections sera formé des deux premiers vecteurs propres, et ainsi de suite.

Théoreme. *Les valeurs propres λ_i et β_i sont identiques.*

Prémultiplions par

$$X^{'}XX^{'}v_i = \beta_i X^{'}v_i$$

Cette équation nous indique que $X^{'}v_i$ est vecteur propre de $X^{'}X$ associé à la valeur propre β_i. Or, les valeurs propres de $X^{'}X$ sont données par λ_i. Donc, nous concluons que $\beta_i = \lambda_i$.

Remarque . Le théorème précédent nous indique qu'il n'est pas nécessaire de rechercher explicitement les valeurs propres et les vecteurs propres de $XX^{'}$. Les valeurs propres sont les mêmes, et les vecteurs propres u_i sont donnés par $X^{'}v_i$ à une constante de normalisation prés. En effet, la norme de $X^{'}v_i$ est:

$$v_i^{'}XX^{'}v_i = \lambda_i$$

donc

$$\frac{1}{\sqrt{\lambda_i}}X^{'}v_i$$

est de norme 1. Par conséquent, nous avons nécessairement

$$\frac{1}{\sqrt{\lambda_i}}X^{'}v_i = u_i$$

et de façon similaire

$$\frac{1}{\sqrt{\lambda_i}}Xu_i = v_i.$$

Par suite, nous aurons les formules de transition sous la forme matricielle, nous pouvons les écrire comme : $V = XU\wedge^{-\frac{1}{2}} \quad U = X^{'}V\wedge^{-\frac{1}{2}}$ où U et V contiennent les vecteurs propres u_i et v_i placés en colonne.

3.9 Reconstruction complète et partielle de la matrice X .

Soit la matrice V ayant les p vecteurs propres de $XX^{'}$ placés en colonne, et U ayant les p vecteurs propres de la matrice $X^{'}X$ placés en colonne. Soit la matrice $\wedge$, une matrice diagonale $p \times p$ ayant les p valeurs propres sur la diagonale. Des formules de transition nous avons

$$XU = V\wedge^{\frac{1}{2}}$$

Nous multiplions cette expression par $U^{'}$ et notant que $UU^{'} = I$, nous obtenons

$$XUU^{'} = X = V\wedge^{\frac{1}{2}}U^{'}$$

Cette dernière expression nous indique que nous pouvons reconstruire la matrice X si nous connaissons les valeurs propres et les vecteurs propres de $XX^{'}$ et $X^{'}X$.

3.10 Coordonnées des individus et des variables sur les vecteurs propres (composantes principales).

Les coordonnées des individus et des variables sont simplement les projections sur les vecteurs propres

- individus : $C_i = XU = V\wedge^{\frac{1}{2}}$
- variables : $C_v = X^{'}V = U\wedge^{\frac{1}{2}}$

3.11 Qualité de la représentation des individus et des variables .

En comparant la projection d'un individu (au carré) avec la distance (au carré) par rapport à l'origine de cet individu, nous obtenons une mesure permettant de juger jusqu'à quel point un individu est prés du vecteur, du plan,... considéré.

- individu i sur vecteur j : $Qlt_i(i,j) = C_i(i,j)^2/l_i^2$
- variable i sur vecteur j : $Qlt_v(i,j) = C_v(i,j)^2/l_i^2$

Remarque l_i^2 se lit sur la diagonale de $XX^{'}$ pour les individus et sur la diagonale de $X^{'}X$ pour les variables.
Dans le cas d'une matrice de corrélation $l_i = 1$.

3.12 Contributions des individus et des variables .

Cette mesure vice à quantifier l'importance de chaque variable et chaque individu dans la définition d'un vecteur propre.

La contribution est donnée par la coordonnée (au carré) d'une observation ou d'une variable sur un vecteur propre divisée par la valeur propre associée à ce vecteur propre. Pour les variables, il suffit de prendre les éléments, au carré, des vecteurs propres.

- individus: $Ctr_i = C_i(.)^2 \wedge^{-1}$
- variables: $Ctr_v = C_v(.)^2 \wedge^{-1} = U(.)^2$

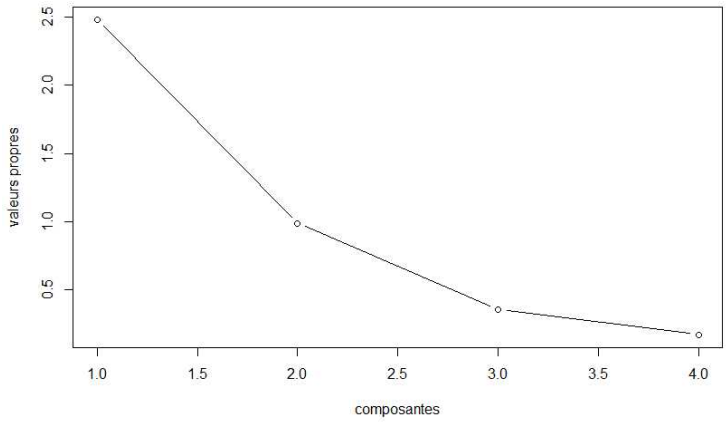
3.13 Exemple numérique complet .

Nous appliquons une ACP sur les donnée du tableau d'arrestation de 4 crimes dans 50 villes aux USA en utilisant le logiciel R, après avoir Charger package "FactoMineR", le code est le suivant:
>library(FactoMineR)
>data(USArrests)
>USArrests
>USArrests.acp<-PCA(USArrests) # les variables sont centrées réduites.
>USArrests.acp$eig calcule les valeurs propres

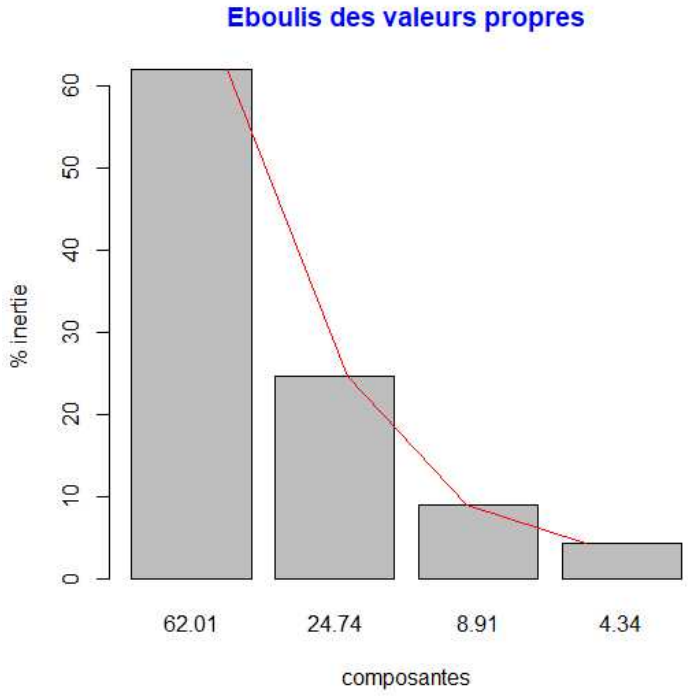
	Valeurs propres	Pourcentage %	Pourcentage cumulé
1	2.4802416	62.006039	62.00604
2	0.9897652	24.744129	86.75017
3	0.3565632	8.914080	95.66425
4	0.1734301	4.335752	100.00000

Interprétation . Chaque ligne du tableau ci-dessus correspond à une variable virtuelle (**les facteurs**) dont la colonne valeurs propres fournit la variance (chaque valeur propre représente la variance du facteur correspondant). La colonne pourcentage de variance, correspond au pourcentage de variance de chaque ligne par rapport au total. La colonne pourcentage cumulé, représente le cumul de ses pourcentage.

>plot(USArrests.acp$eig[,1]type="b",ylab="valeurs propres",xlab="composantes",Iwd=
représentation graphique



>barplot(USArrests.acp$eig[,2],ylab="% inertie",xlab="composantes", names.arg=round(USArrests.acp$eig[,2],2),main="Eboulis des valeurs propres",col.main="blue")
>lines(USArrests.acp$eig[,2],col="red") # visualiser les valeurs propres afin de montrer le pourcentage variances expliquées pour chaque axe principale.

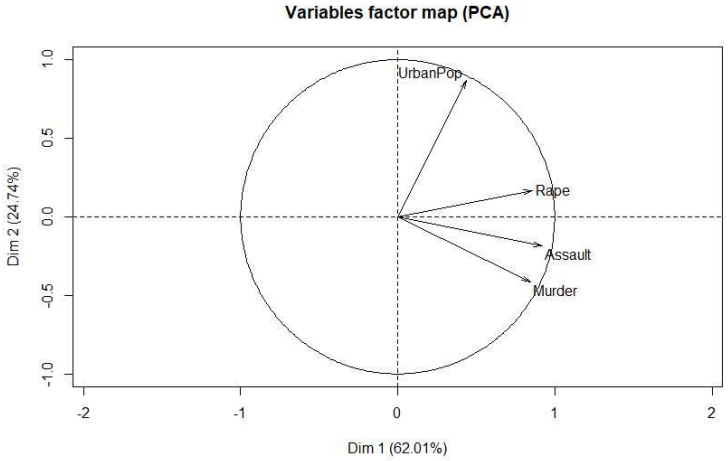


L'inertie totale est répartie selon 4 valeurs propres, nous n'allons considérer que 2 valeurs propres car l'inertie cumulé sera 86.75
Avec FactoMineR, nous obtenons tous les tableaux de résultats sur les individus et les variables, nous avons
>USArrest.acp$var

Var	F1	COS2	CTR	F2	COS2	CTR
Murder	0.84	0.71	28.72	−0.42	0.17	17.49
Assault	0.92	0.84	34.01	−0.19	0.03	3.53
UrbanPop	0.44	0.19	7.74	0.87	0.75	76.18
Rape	0.86	0.73	29.53	0.17	0.03	2.80

Pour le tableau associé aux individus, nous avons pas pu le mettre car les calculs sont plus longs, nous avons donné juste la commande qui nous permet de calculer les coordonnées, la coorélation et la contribution.

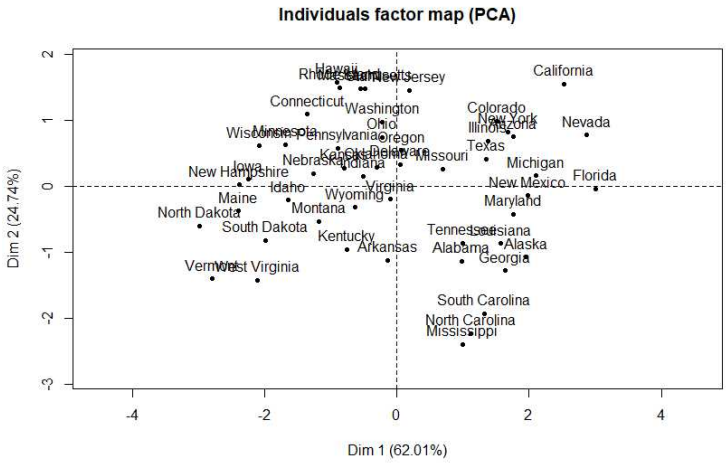
>USArrests.acp$ind
Pour la représentation des variables et des individus , le code est
>plot(USArrests.acp,choix="var",autolab="yes") # pour les variaibles



Interprétation .nous voyons que le premier facteur est correlé positivement, avec chacune des 4 variables initiales: plus un criminel a des crimes grave, plus il a un score élevé sur l'axe 1, réciproquement, plus ses crimes sont pas grave, plus son score est négatif; l'axe 1 représente donc le résultat globale des crimes. En ce qui concerne l'axe 2, il oppose d'une part UrbanPop et Rape (corrélations positives), d'autre part, Assault et Murder (corrélations négatives).

>**plot(USArrests.acp,choix="ind",autolab="yes")** #pour les individus

Les composantes principales sont les nouvelles variables sur lesquelles nous nous appuyons pour visualiser les individus. Cette représentation approchée des individus sur un espace de dimension 2.



Chapitre 4

Analyse factorielle des correspondances

L'AFC est une analyse destinée au traitement des tableaux de données où les valeurs sont positives et homogène comme les tableaux de contingence. Le but de l'AFC est de lire l'information contenue dans un espace multidimenssionnel par une réduction de la dimension de cet espace tout en concervant un maximum de l'information contenu dans l'espace de départ.

4.1 Tableau de contingence .

Le tableau de contingence est un moyen particulier de représenter simultanément deux caractères observès sur une même population, s'ils sont discrets ou bien continus et regroupés en classes. Les deux caractères sont V et W. Les modalités ou classes de V seront notées $v_1, ..., v_r$, celles de W sont notées $w_1, ..., w_s$, l'effectif total sera noté N. Nous notons

- $N = (n_{hk})_{h=1,...,r,k=1,...,s}$ l'effectif conjoint de V_h et W_k: c'est le nombre d'individus pour lesquels V prend la valeur v_h et W la valeur w_k,
- $n_{h\bullet} = \sum_{k=1}^{s} n_{hk}$ l'effectif marginal de v_h: c'est le nombre d'individus pour lesquels V prend la valeur v_h,
- $n_{\bullet k} = \sum_{h=1}^{r} n_{hk}$ l'effectif marginal de w_k: c'est le nombre d'individus pour lesquels W prend la valeur w_k.

Nous représentons ces valeurs dans un tableau de contingence :

$V \setminus W$	w_1	...	w_k	...	w_s	total
v_1	n_{11}	...	n_{1k}	...	n_{1s}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
v_h	n_{h1}	...	n_{hk}	...	n_{hs}	$n_{h\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
v_r	n_{r1}	...	n_{rk}	...	n_{rs}	$n_{r\bullet}$
total	$n_{\bullet 1}$	...	$n_{\bullet k}$	...	$n_{\bullet s}$	N

Nous définissons les marges en fréquence avec $f_{hk} = n_{hk}/N$

$$f_{h\bullet} = \sum_{k=1}^{s} f_{hk}, f_{\bullet k} = \sum_{h=1}^{r} f_{hk}, f_{\bullet\bullet} = \sum_{h,k} f_{hk} = 1$$

Poids des points ligne et points colonne .

1. Chaque point ligne est doté d'un poid relatant l'importance de la modalité h de V $f_{h\bullet} = \dfrac{n_{h\bullet}}{N}$
2. Chaque point colonne est doté d'un poid relatant l'importance de la modalité k de W $f_{\bullet k} = \dfrac{n_{\bullet k}}{N}$

4.2 Modèle d'indépendance .

4.2.1 Test de chi 2 .

Comme en ACP, nous nous intéressons alors aux direction de "plus grande dispersion" de chacun de ces nuages de points, mais nous utilisons la distance de χ^2 entre ces deux variables (à la place de la distance euclidienne). Cette distance permet de comparer l'effectif de chacune des cellules du tableau de contingence à la valeur qu'elle aurait si les deux variables etaient indépendantes.
Notons E_{hk} l'effectif attendu sous l'hypothèse d'indépendance; par définition

$$E_{hk} = \frac{n_{h\bullet} n_{\bullet k}}{N}$$

Et la distance du χ^2 est définie par

$$d^2_{\chi^2}(N, E) = \sum_{h=1}^{r} \sum_{k=1}^{s} \frac{(n_{hk} - E_{hk})^2}{E_{hk}}$$

Plus la distance $d^2_{\chi^2}(N,E)$ est grande, plus le tableau observé est éloigné du tableau attendu sous l'hypothèse d'indépendance.
Sous l'hypothèse d'indépendance des deux variables, la statistique $d^2_{\chi^2}$ suit une loi du χ^2 à $(r-1)(s-1)$ degré de liberté. Cette loi sert par exemple, à définir une règle de décision de type: nous concluons que les variables sont indépendantes avec un risque α si $d^2_{\chi^2}(N,E) < F^{-1}_{(r-1)(s-1)}(1-\alpha)$ avec F la fonction de répartition de la loi du χ^2 à $(r-1)(s-1)$ degré de liberté.

4.3 La transformation initiale des données .

L'espace $\mathbb{R}^s$ des variables (modalités colonnes) dans lequel nous pouvons représenter le nuage des r points $\ll$ individus $\gg$ (modalités lignes). Chaque individu a pour coordonnée $X_{hk} = \dfrac{f_{hk}}{f_{h\bullet}}$ et dans cet espace nous utilisons le tableau des profils lignes.
L'espace $\mathbb{R}^r$ individus dans lequel nous pouvons représenter le nuage des s points $\ll$ variables $\gg$. Chaque variable a pour coordonnée $Y_{hk} = \dfrac{f_{hk}}{f_{\bullet k}}$ et dans cet espace nous utilisons le tableau des profils colonnes.
La distance entre deux profils lignes et profils colonnes . Nous calculons la distance euclidienne entre deux point h et $h^{'}$ dans l 'espace $\mathbb{R}^r$:

$$d^2_{\chi^2}(h,h^{'}) = \sum_{k=1}^{s}\left(\frac{f_{hk}}{f_{h\bullet}} - \frac{f_{h^{'}k}}{f_{h^{'}\bullet}}\right)^2$$

En AFC est contrairement à l'ACP, nous n'utilisons pas cette distance euclidienne. Plus précisément, nous l'utilisons mais aprés avoir effectué une transformation préalable des coordonnées des points du nuages. Dans l'espace $\mathbb{R}^r$ cette transformation s'écrit: $X_{hk} = \dfrac{1}{\sqrt{f_{\bullet k}}}\dfrac{f_{hk}}{f_{h\bullet}}$
En définitive, dans l'espace $\mathbb{R}^r$ nous calculons la distance entre deux points h et $h^{'}$ par la formule
$d^2_{\chi^2}(h,h^{'}) = \sum_{k=1}^{s}\left(\dfrac{1}{\sqrt{f_{\bullet k}}}\dfrac{f_{hk}}{f_{h\bullet}} - \dfrac{1}{\sqrt{f_{\bullet k}}}\dfrac{f_{h^{'}k}}{f_{h^{'}\bullet}}\right)^2 = \sum_{k=1}^{s}\dfrac{1}{f_{\bullet k}}\left(\dfrac{f_{hk}}{f_{h\bullet}} - \dfrac{f_{h^{'}k}}{f_{h^{'}\bullet}}\right)^2$
Nous procèdons de façon équivalente pour l'espace $\mathbb{R}^s$, nous considèrons maintenant deux points du nuage k et $k^{'}$, la transformation $Y_{hk} = \dfrac{1}{\sqrt{f_{h\bullet}}}\dfrac{f_{hk}}{f_{\bullet k}}$ conduit à la distance
$d^2(k,k^{'}) = \sum_{h=1}^{r}\left(\dfrac{1}{\sqrt{f_{h\bullet}}}\dfrac{f_{hk}}{f_{\bullet k}} - \dfrac{1}{\sqrt{f_{h\bullet}}}\dfrac{f_{hk^{'}}}{f_{\bullet k^{'}}}\right)^2 = \sum_{h=1}^{r}\dfrac{1}{f_{h\bullet}}\left(\dfrac{f_{hk}}{f_{\bullet k}} - \dfrac{f_{hk^{'}}}{f_{\bullet k^{'}}}\right)^2$

4.3.1 AFC et indépendance .

L'analyse d'un tableau de contingence doit donc se faire en référence à la situation de d'indépendance. C'est ce que fait l'AFC en écrivant le modèle

d'indépendance sous la forme suivante

$$\forall h = 1, ..., r, \forall k = 1, ..., s, \frac{f_{hk}}{f_{h\bullet}} = f_{\bullet k}$$

La quantité $\frac{f_{hk}}{f_{h\bullet}}$ est la probabilité conditionnelle de posséder la modalité k de la variable Y sachant que nous possèdons la modalité h de la variable X. De façon symétrique, nous pouvons écrire

$$\forall h = 1, ..., r, \forall k = 1, ..., s, \frac{f_{hk}}{f_{\bullet k}} = f_{h\bullet}$$

Définition 13.

- *L'ensemble de probabilité* $\{f_{hk}/f_{h\bullet}, k = 1, ..., s\}$ *est appelée profil ligne.*
- *L'ensemble de probabilité* $\{f_{hk}/f_{\bullet k}, h = 1, ..., r\}$ *est appelée profil colonne .*
- $\{f_{h\bullet}; k = 1, ..., s\}$ *(resp.*$\{f_{\bullet k}; h = 1, ..., r\}$*) est le profil moyen correspondant au profil ligne (resp.colonne).*

Remarque . Si nous avons indépendance, le profile ligne d'une part et colonne d'autre part est égal au profil moyen correspondant.

4.4 Inertie total .

Nous définissons l inertie total par :

$$\begin{aligned} I_T &= f_{h\bullet} \sum_{h=1}^{r} d^2_{\chi_2}(h, G_h) \\ &= f_{h\bullet} \sum_{h=1}^{r} \sum_{k=1}^{s} \frac{1}{f_{\bullet k}} \left(\frac{f_{hk}}{f_{h\bullet}} - f_{\bullet k} \right)^2 \\ &= \sum_{h=1}^{r} \sum_{k=1}^{s} \frac{(f_{hk} - f_{h\bullet} f_{\bullet k})^2}{f_{h\bullet} f_{\bullet k}} \end{aligned}$$

4.5 L'AFC proprement dite .

Pour étudier les lignes, nous pouvons réaliser une ACP de la matrice A (telle que $a_{hk} = n_{hk}/n_{h\bullet}$) puis de représenter les modalités de la première

variable. En raison de changement de métrique, nous introduisons la matrice $M = D_C^{-1}$ avec $D_c = diag(n_{\bullet 1}, ..., n_{\bullet s})$ et nous considèrons la matrice de poids $D = D_L^{-1}$ avec $D_L = (n_{1\bullet}, ..., n_{r\bullet})$. nous remarquons que $A = D_L^{-1}N$. De façon symétrique, nous pouvons définir $B = ND_c^{-1}$.

4.6 Détermination des composantes principales dans $\mathbb{R}^s$.

4.6.1 Caractéristiques des variables et construction de la matrice d'information .

Comme pour l'ACP nous nous plaçons dans l'espace des variables, nous utilisons donc la matrice des profils lignes. nous venons de voir que dans cet espace les r points du nuage ont pour coordonnées $X_{hk} = \frac{1}{\sqrt{f_{\bullet k}}}\frac{f_{hk}}{f_{h\bullet}}$
Nous pouvons calculer les caractéristiques de ces variables (notées X_k pour $k = 1, .., s$) c'est-à-dire la moyenne et la covariance.
La moyenne .

$$\begin{aligned}
\overline{X}_k &= \sum_h f_{h\bullet} X_{hk} \\
\overline{X}_k &= \sum_h f_{h\bullet} \frac{1}{\sqrt{f_{\bullet k}}} \frac{f_{hk}}{f_{h\bullet}} \\
&= \sum_h \frac{f_{hk}}{\sqrt{f_{\bullet k}}} = \frac{1}{\sqrt{f_{\bullet k}}} \sum_h f_{hk} = \frac{1}{\sqrt{f_{\bullet k}}} f_{\bullet k} = \sqrt{f_{\bullet k}}
\end{aligned}$$

La covariance . La covariance entre deux variables X_k et $X_{k'}$ est

$$\begin{aligned}
cov(X_k, X_{k'}) &= V_{kk'} = \sum_h f_{h\bullet} \left[\left(\frac{1}{\sqrt{f_{\bullet k}}} \frac{f_{hk}}{f_{h\bullet}} - \sqrt{f_{\bullet k}} \right) \left(\frac{1}{\sqrt{f_{\bullet k'}}} \frac{f_{hk'}}{f_{h\bullet}} - \sqrt{f_{\bullet k'}} \right) \right] \\
cov(X_k, X_{k'}) &= \sum_h \frac{f_{hk} f_{hk'}}{\sqrt{f_{\bullet k}} \sqrt{f_{\bullet k'}} f_{h\bullet}} - \sqrt{f_{\bullet k}} \sqrt{f_{\bullet k'}}
\end{aligned}$$

La matrice d'information (matrice des variances des covariances des variables) . En faisant varier k et $k^{'}$ de 1 à s nous construisons alors la matrice $V(s, s)$ des variances covariances des variables. C'est la matrice d'information des variables et par analogie avec l'ACP, l'étape suivante de L'AFC sera la diagonalisation de cette matrice .

4.6.2 Diagonalisation de la matrice des variances covariances ou de la matrice d'inertie .

La matrice V précédente permet de calculer par les valeurs propres et les vecteurs propres normés, la matrice du changement de base. En AFC nous n'utilisons pas toujours la matrice V mais une matrice plus simple appelée la matrice d'inertie. Considérons la matrice $V(s,s)$ de terme général $V_{kk'}$, nous pouvons démontrer que le premier vecteur propre u_0 issu de cette matrice V a pour coordonnées :

$$u_0 = \begin{bmatrix} u_{01} \\ \cdots \\ u_{0k} \\ \cdots \\ u_{0s} \end{bmatrix} = \begin{bmatrix} \sqrt{f_{\bullet 1}} \\ \cdots \\ \sqrt{f_{\bullet h}} \\ \cdots \\ \sqrt{f_{\bullet s}} \end{bmatrix}$$

Et qu'il est associé à la première valeur propre $\lambda_0 = 0$ de V, nous savons que tous les vecteurs propres sont orthogonaux deux à deux. Donc le produit scalaire d'un vecteur propre u_p quelconque de V avec u_0 est égal à zéro, ce qui s'écrit : $u_p^{'} \times u_0 = 0 \Leftrightarrow$

$$[\ u_{p1} \ \ .. \ \ u_{pk} \ \ .. \ \ u_{ps}\ ] \begin{bmatrix} u_{01} \\ \cdots \\ u_{0k} \\ \cdots \\ u_{0s} \end{bmatrix} = u_{p1} \times u_{01} + .. + u_{pk} \times u_{0k} + .. + u_{ps} \times u_{0s} = 0$$

$\Leftrightarrow \sum_{k=1}^{s} u_{pk} u_{0k} = 0 \Leftrightarrow \sum_{k=1}^{s} u_{pk} \sqrt{f_{\bullet k}} = 0 \rightarrow$ (relation 1)

Ecrivons que u_p est le vecteur propre associé à la valeur propre λ_p de la matrice V

$$V u_p = \lambda_p u_p \qquad [7]$$

Soit encore $\Leftrightarrow [V_{kk'}] \begin{bmatrix} u_{p1} \\ \cdots \\ u_{pk} \\ \cdots \\ u_{ps} \end{bmatrix} = \lambda_p \begin{bmatrix} u_{p1} \\ \cdots \\ u_{pk} \\ \cdots \\ u_{ps} \end{bmatrix}$

pour le j^{me} terme $V_{k1}u_{p1} + ... + V_{kk'}u_{pk'} + ... + V_{ks}u_{ps} = \lambda_p u_{pk}$.

$\Leftrightarrow \sum_{k'}^{s} V_{kk'} u_{pk'} = \lambda_p u_{pk}$, nous remplaçons $V_{kk'}$ par sa valeur

$$\begin{aligned} \lambda_p u_{pk} &= \sum_{k'}^{s} \left(\sum_{h} \frac{f_{hk} f_{hk'}}{\sqrt{f_{\bullet k}} \sqrt{f_{\bullet k'}} f_{h\bullet}} - \sqrt{f_{\bullet k}} \sqrt{f_{\bullet k'}} \right) u_{pk'} \\ &= \sum_{k'} \sum_{h} \frac{f_{hk} f_{hk'}}{\sqrt{f_{\bullet k}} \sqrt{f_{\bullet k'}} f_{h\bullet}} u_{pk'} - \sum_{k'} \sqrt{f_{\bullet k}} \sqrt{f_{\bullet k'}} u_{pk'} \\ &= \sum_{k'} \sum_{h} \frac{f_{hk} f_{hk'}}{\sqrt{f_{\bullet k}} \sqrt{f_{\bullet k'}} f_{h\bullet}} u_{pk'} - \sqrt{f_{\bullet k}} \sum_{k'} \sqrt{f_{\bullet k'}} u_{pk'} \end{aligned}$$

Or d'aprés la relation 1 on a $\sum_{k'} \sqrt{f_{\bullet k'}} u_{pk'} = 0$

D'où $\lambda_p u_{pk} = \sum_{k'} \left(\sum_{h} \frac{f_{hk} f_{hk'}}{\sqrt{f_{\bullet k}} \sqrt{f_{\bullet k'}} f_{h\bullet}} u_{pk'} \right)$ (Relation 2)

Nous posons $\sum_{h} \frac{f_{hk} f_{hk'}}{\sqrt{f_{\bullet k}} \sqrt{f_{\bullet k'}} f_{h\bullet}} = S_{kk'}$, et la relation 2 s'écrit:

$$\lambda_p u_{pk} = \sum_{k'} S_{kk'} u_{pk'} = \sum_{k'}^{s} V_{kk'} u_{pk'}$$

Appelons S la matrice de terme $S_{kk'}$ nous avons alors:

$$\lambda_p u_p = V u_p = S u_p$$

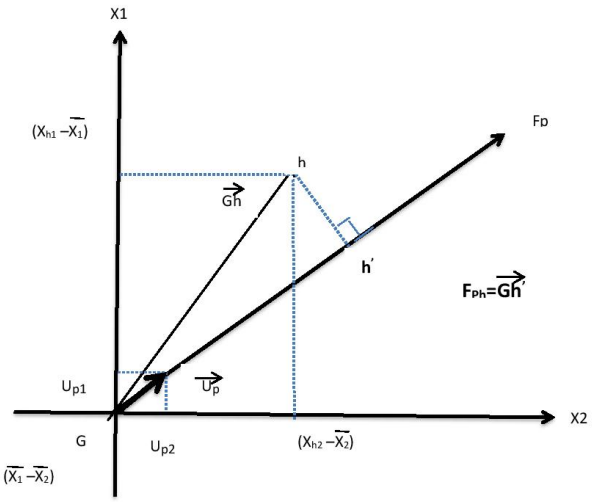
La matrice S porte le nom de matrice d'inertie. Nous pouvons constater que les vecteurs propres de la matrice V sont identiques à ceux de S. Il est donc indifférent de diagonaliser la matrice S ou V.
Le premier vecteur propre associé à cette première valeur propre définit un axe principal pour lequel les projections des individus et des variables possèdent une variance (dispersion) nulle. Ce qui signifie que toutes les projections possèdent les mêmes coordonnées.

4.6.3 Le choix du nombre de composantes principales

Nous avons déja vu que $I(h, G_I) = f_{h\bullet} \sum_{h=1}^{r} d^2_{\chi_2}(h, G_h)$ et
$d^2_{\chi_2}(h, G_h) = \sum_{k=1}^{s} \left(\frac{f_{hk}}{f_{h\bullet}\sqrt{f_{\bullet k}}} - \sqrt{f_{\bullet k}} \right)^2 = \sum_k \left(X_{hk} - \overline{X_k} \right)^2$
Nous constatons que ce moment d'inertie est la variance multidimensionnelle dont nous savons en ACP qu'elle est aussi donnée par la trace de la matrice d'information S ou V. En définitive nous pouvons écrire que $I = \sum_k S_{kk} = tr(S) = \sum_k V_{kk}$.
Or $tr(S) = \sum_k \lambda_k$

4.7 Les coordonnées des projections des individus et des variables sur les axes principaux

Contrairement à l'ACP, en AFC les projections sur les axes principaux du nuage des individus et des variables s'effectuent sur un même graphique. Nous parlons de projection simultanée. Pour rappeler comment nous réalisons la projection d'un point sur un axe factoriel, considérons par exemple, un individu h et un axe principal noté F_p et plaçons nous dans un espace à deux dimensions de deux variables X_1 et X_2.



La projection orthogonale du point h sur l'axe F_p est donnée par le produit scalaire

$$\overrightarrow{Gh'} = \overrightarrow{Gh} \times \overrightarrow{u_p} = \overrightarrow{u_p} \times \overrightarrow{Gh} \qquad [7]$$

Or ces vecteurs ont les coordonnées suivantes : $\overrightarrow{u_p}=\begin{pmatrix} u_{p1} \\ u_{p2} \end{pmatrix}$ et $\overrightarrow{Gh}=\begin{pmatrix} X_{h1}-\overline{X}_1 \\ X_{h2}-\overline{X}_2 \end{pmatrix}$ donc

$\overrightarrow{Gh'} = \overrightarrow{u_p} \times \overrightarrow{Gh} = \begin{bmatrix} u_{p1} & u_{p2} \end{bmatrix} \begin{bmatrix} X_{h1}-\overline{X}_1 \\ X_{h2}-\overline{X}_2 \end{bmatrix} = u_{p1}(X_{h1}-\overline{X}_1) + u_{p2}(X_{h2}-\overline{X}_2)$

D'où $\overrightarrow{Gh'} = F_{ph} = \sum_{k=1}^{2} u_{pk}(X_{hk}-\overline{X}_k)$

En généralisant ce résultat à l'espace complet $\mathbb{R}^s$ on a

$F_{ph} = \sum_{k=1}^{s} u_{pk}\left(\frac{f_{hk}}{f_{h\bullet}\sqrt{f_{\bullet k}}} - \sqrt{f_{\bullet k}}\right) = \sum_k u_{pk}\frac{f_{hk}}{f_{h\bullet}\sqrt{f_{\bullet k}}} - \sum_k u_{pk}\sqrt{f_{\bullet k}}$

D'aprés la relation (1) $\sum_k u_{pk}\sqrt{f_{\bullet k}} = 0$

Posons maintenant $a_{pk} = \dfrac{u_{pk}}{\sqrt{f_{\bullet k}}}$:

$$\boxed{F_{ph} = \sum_{k=1}^{s} u_{pk}\frac{f_{hk}}{f_{h\bullet}\sqrt{f_{\bullet k}}} = \sum_k \frac{f_{hk}}{f_{h\bullet}}a_{pk}} \quad (A)$$

Cette relation permet de vérifier que les coordonnées de tous les individus sur l'axe principale qui a pour vecteur unitaire $u_{0k}=\sqrt{f_{\bullet k}}$ sont égales à

1. En effet, pour $p = 0$ nous avons :

$$\begin{aligned} F_{0h} &= \sum_{k=1}^{s} u_{0k} \frac{f_{hk}}{f_{h\bullet}\sqrt{f_{\bullet k}}} \\ &= \sum_{k=1}^{s} \sqrt{f_{\bullet k}} \frac{f_{hk}}{f_{h\bullet}\sqrt{f_{\bullet k}}} \\ &= \frac{1}{f_{h\bullet}} \sum_{k} f_{hk} = \frac{f_{h\bullet}}{f_{h\bullet}} = 1 \end{aligned}$$

donc F_{0h}=1 quel que soit h.
Nous pourrions vérifier avec les formules de transition que a_{pk} correspond à la projection orthogonale de la variable k sur l'axe principale p noté a_p, de ce fait, la formule A: $F_{ph} = \sum_k \frac{f_{hk}}{f_{h\bullet}} a_{pk}$ n'est autre que le calcul du centre de gravité (de la moyenne pondérée) des coordonnées des projections des k variables, d'où la propriété barycentrique de l'AFC : Les coordonnées des projections orthogonales de chaque point h sont le barycentre (moyenne pondérée) des coordonnées des projections des points k. Et réciproquement, les coordonnées des projections de chaque point k sont le barycentre des projections des points h.

Cette propriété découle des Formules de Transition entre le tableau des profils lignes de l'espace $\mathbb{R}^s$ et celuis des profils colonnes de l'espace $\mathbb{R}^r$. Cette propriété barycentrique permet donc d'écrire que

$$\boxed{a_{pk} = \sum_{h=1}^{s} \frac{f_{hk}}{f_{\bullet k}} F_{ph}} \ (B), \text{ ou encore } \boxed{F_{ph} = \sum_{k=1}^{r} \frac{f_{hk}}{f_{h\bullet}} a_{pk}} \ (A)$$

Les formules de transitions montrent alors que la réalisation simultanée des écritures A et B n'est pas possible. Pour que cela le soit, il faut introduire dans les formules précédentes le paramètre $\frac{1}{\sqrt{\lambda_p}}$.
Les formules des projections simultanées des variables et des individus s'écrivent alors
$\hat{F_{ph}}$=F_{ph}=$\sum_k u_{pk} \frac{f_{hk}}{f_{h\bullet}\sqrt{f_{\bullet k}}}$, ou encore $\hat{F_{ph}}$=$\frac{1}{\sqrt{\lambda_p}} \sum_{k=1}^{r} \frac{f_{hk}}{f_{h\bullet}} \hat{a_{pk}}$
et $\hat{a_{pk}}$=$\sqrt{\lambda_p} a_{pk}$, ou encore $\hat{a_{pk}}$=$\frac{1}{\sqrt{\lambda_p}} \sum_{h=1}^{r} \frac{f_{hk}}{f_{\bullet k}} \hat{F_{ph}}$
Remarque . $\wedge$ signifie la valeur calculée.
Par exemple $\hat{F_{ph}}$ (le calcul de la projection de h sur F_p) peut être calculé en utilisant la formule F_{ph} de la relation A ou bien en utilisant la formule A où nous connaissons $\hat{a_{pk}}$.

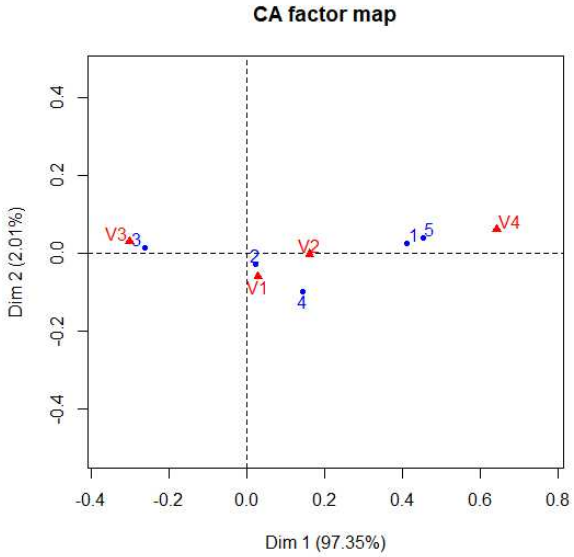
Exemple .
Nous avons un tableau des étudiants de première année et choix d'un secteur disciplinaire.

	Droit	Science	Médecine	IUT	TOTAL
Exp.agri.	80	99	65	58	302
Patron	168	137	208	62	575
Cadre sup.	470	400	876	79	1825
Employé	145	133	135	54	467
Ouvrier	166	193	127	129	615
TOTAL	1029	962	1411	382	3784

Nous faisons entrer les données de ce tableau dans le logiciel R:
>CSD<-c("Droit","Science","Médecine","IUT")
> OSE<-c("Exp.agr","Patron","Cadre sup","Employé","Ouvrier")
> donnée.AFC<-matrix(c(80,99,65,58,168,137,208,62,470,400,876,79,145,133,135,
54,166,193,127,129),nrow=5 ,byrow=TRUE)
> donnée.AFC

Nous allons calculer la χ^2_{calcul} en utilisant la commande **chisq.test(donnée.AFC)** du logiciel R où nous trouvons 320.27
Or pour un risuqe $\alpha = 0.05$ on a $\chi^2 = 13.84$, donc $\chi^2_{calcul} > \chi^2_{12}$
Pour appliquer l'AFC nous utilisons la commande **library(FactoMineR)**
>CA(donnée.AFC)$eig. # Pour calculer les valeurs propres.

Nous obtenons le graphique suivant. Nous observons que toutes les modalités sont concentrées autour du premier axe. Ceci signifie que nous avons essentiellement une seule variable latente (ou facteur) structurante.



Les résultats pour les profils colonnes:
>CA(donnée.AFC)colcoord

.	Dim 1	Dim 2	Dim 3
V1	0.02798724	−0.060669159	0.016544781
V2	0.16046170	−0.002734195	−0.037582581
V3	−0.30312512	0.029661814	0.005200252
V4	0.64017413	0.060748795	0.030869913

> CA(donnée.AFC)colcos2

.	Dim 1	Dim2	Dim 3
V1	0.1653282	0.7768956014	0.0577761737
V2	0.9477351	0.0002751713	0.0519897127

V3	0.9902269	0.0094817015	0.0002914336
V4	0.9887968	0.0089040202	0.0022992257

>CA(donnée.AFC)colcontrib

.	Dim 1	Dim 2	Dim 3
V1	0.2585182	58.7585597	13.78948
V2	7.9446214	0.1115716	66.52097
V3	41.5839983	19.2593782	1.86804
V4	50.2128622	21.8704905	17.82151

Interprétation .

Contribution du profil colonne . $CTR(k)=\frac{f_{\bullet k}a_{pk}^2}{\lambda_p}$ tel que $\lambda_p=\sum_{k=1}^{s} f_{\bullet k}a_{pk}^2$

Nous appelons le vecteur de profil colonne $B=$ (f_{1k} ... f_{hk} ... f_{rk})

Qualité de représentation . $qlt(k)=\frac{a_{pk}^2}{\|B-G_k\|^2}$

Lorsque l'angle est proche de 0, c'est-à-dire que le cosinus est proche de 1, l'individu est bien représenté. Dans le cas inverse, l'angle est proche de 90° et le cosinus est proche de 0.

Les résultatus pour les profils lignes:

> CA(donnée.AFC)rowcoord

.	Dim 1	Dim 2	Dim 3
1	0.41011544	0.02625317	−0.038283778
2	0.02015079	−0.02658535	0.046880589
3	−0.26271704	0.01559580	−0.006198846
4	0.14209032	−0.09732566	−0.021242138
5	0.45148105	0.03958841	0.009493228

> CA(donnée.AFC)rowcos2

.	Dim 1	Dim 2	Dim 3
1	0.9873503	0.004045968	0.0086037660
2	0.1226519	0.213488665	0.6638594687
3	0.9959358	0.003509700	0.0005544682
4	0.6704594	0.314556176	0.0149844193
5	0.9919347	0.007626762	0.0004385628

> CA(donnée.AFC)rowcontrib

.	Dim 1	Dim 2	Dim 3
1	16.29200620	3.229165	21.669399
2	0.07488716	6.304816	61.867745
3	40.40124242	6.886489	3.433167
4	3.02413561	68.626434	10.316292
5	40.20772861	14.953096	2.713396

Interprétation .

Contribution du profil ligne . $CTR(h)=\frac{f_{h\bullet}F_{ph}^2}{\lambda_p}$ tel que $\lambda_p=\sum_{h=1}^{r} f_{h\bullet}F_{ph}^2$

Nous appelons le vecteur de profil ligne $A=$ (f_{h1} ... f_{hk} ... f_{hs})

Qualité de représentation . $qlt(h)=\frac{F_{ph}^2}{\|A-G_h\|^2}$

Conclusion générale

Les analyses multivariées permettent de prendre en compte les facteurs de confusion, en ajustant sur ces facteurs. Elles sont donc recommandées lorsqu'on cherche à établir un lien statistique entre plusieurs variables. Les analyses multivariées font appel à des méthodes statistiques plus sophistiquées que les analyses univariées, et sont rarement disponibles dans les logiciels à destination des non statisticiens. Les deux méthodes factorielles ou descriptives qui ont pour but de proposer un nouveau système de représentation des variables latentes formées à partir de combinaisons linéaires des variables prédictives qui permettent de discerner le plus possible les groupes d'individus.
Dans ce travail, nous avons étudié et donné quelques résultats élémentaires pour ces deux méthodes. Des exemples d'applications ont étaient traités pour chacune des deux méthodes en utilisant le langage R pour bien interpréter des données réelles.

Bibliographie

[1] Jean-Baptiste Bardet , 10 Février 2006, *Théorme de Cochran et applications en statistiques*

[2] Bounkhala Asma , Septembre 2017,spécialité statistiques et probabilité approfondie, *Méthodes ACP et AFC en statistiques et leurs applications.*

[3] Professeur Michel Carbon,Département de Mathématique et Statistique Université de Laval, Hiver 2015,*Cours D'analyse de la variance.*

[4] C.DUBY , S.ROBIN, *Analyse en composantes principales ..*

[5] Ane B Dufour, Octobre 2013. *Analyse en composantes principales ..*

[6] Arnaud Guyader , *Espérance conditionnelle et chaînes de Markov ..*

[7] L'Analyse Factorielle des Correspondance, http://www.foad-mooc.auf.org/IMG/pdf/M05-3.pdf

[8] V.Monbet,Master 1 -2013-2014 *Analyse des données ,Analyse Statistique et économétrie..*

[9] Tableau de contingence, https://mistis.inrialpes.fr/software/SMEL/cours/sd/node16.html

Résumé

L'objectif général des méthodes d'analyse factorielle est la recherche de facteurs permettant de résumer les données, ces méthodes visent à réduire la dimension des données (le nombre de variables) en conservant au mieux l'information utile. Dans le cadre de ce travail, nous allons nous limiter à deux de ces méthodes : L'analyse en composantes principales (en sigle ACP) et l'analyse factorielle des correspondants (AFC), en donnant un exemple d'application pour chacune d'elle.

Abstract

The général objectif of factor analys is methods is to identify factors that can be used to summarize the data. These methods aim to reduce the size of the data (the number of variables) by keeping as much useful information as possible. In this work, we will limit our selves to two of these methods : PCA and AFC, giving an example of an application for each of them.