

# **CARACTERISATION DES ACCESSIONS DE SAFOUTIERS**

Par :

**Mbogning Tchinda Cyprien**

Maitre és-Sciences Mathématiques

Sous la direction de

**Ndoumbe Nkeng Michel**

Docteur

et de

**Kouediekong Lazare**

---

# Dédicaces

---

A mes Parents,

Mme TCHINDA née MAFOU Madeleine et Mr TCHINDA Jean Bart

Maman, Papa

Vous m'avez appris le sens de la famille et des amis.

Vous m'avez appris le sens du travail et de la persévérance.

Vous m'avez toujours donné les conditions d'études dans la limite de vos moyens et un cadre familiale favorable.

Merci maman, merci papa pour tout, voyez en ce mémoire un des fruits de tous les efforts et sacrifices que vous ne cessez de faire pour nous.

A tous mes petits frères et soeurs,

Mr FOPA TCHINDA Elie, Mr SOP TCHINDA Giscard, Mle FOUTSOP TCHINDA Mariette,  
Mle MAFOI TCHINDA Elodie et Mle KENNE TCHINDA Mireille

A ma soeur et son époux,

Mme KOUONKANG née MALIAH TCHINDA Sylvienne et Mr KOUONKANG Rigobert

Ce travail est le fruit de vos efforts.

Sans oublier mes grand-parents pour toutes leurs bénédictions.

---

# Remerciements

---

**Au Professeur Henri Gwet**

pour l'intérêt qu'il porte à mes travaux.

**Au Docteur Doumbe Nkeng**

Pour son aide académique et logistique.

**A Mr Kouediekong Lazare**

Pour son aide précieuse et sa disponibilité tout au long de ces travaux.

**Au Docteur Eugène Patrice Ndong Nguéma**

pour tous ses enseignements et surtout pour le fait de m'avoir rendu passionné pour la statistique.

**A Takam Patrice**

Pour ses précieux conseils et sa disponibilité.

**A tout le personnel du PRP Fruit de l'IRAD Nkolbisson.**

Pour l'accueil qu'ils m'ont réservé pendant mon séjour dans leur structure.

**A tout le corps enseignant du Master.**

Pour tous les enseignements sans lesquels ces travaux n'auraient pas pu être réalisés.

**A tous mes camarades du Master.**

Avec qui tout au long de l'année on a eu à travailler dur pour arriver à nos fins

**A ma famille.**

Pour le soutien sans faille et la patience qu'elle a su me montrer.

**A tous mes potes.**

**Je remercie toutes les personnes dont j'aurai oublié le nom et qui auraient contribué d'une manière ou d'une autre à la réalisation de ce travail.**

---

# Introduction

---

## Contexte

Dans toute la région de l'Afrique Centrale et du Golfe de Guinée, le safoutier (*Dacryodes edulis*) est l'espèce fruitière traditionnelle la plus cultivée. Ses fruits (les safous) sont riches en acides gras (jusqu'à 65% de la pulpe) et en acides aminés, d'où l'appellation de " bush butter " en anglais. Ces fruits présentent par conséquent un intérêt alimentaire certain et constituent une source de revenus non négligeables pour les paysans.

Malgré son importance sur le double plan économique et alimentaire, le safou n'a pas encore fait l'objet d'études nécessaires à son amélioration et à sa valorisation. Ce dernier est sous exploité dans les pays en développement par manque d'informations scientifiques.

## Enjeu

Le travail que nous réalisons permettra aux chercheurs dans la filière *safou* d'avoir de meilleures connaissances sur les différents écotypes et surtout d'effectuer des sélections variétales ; ce qui n'était pas fréquent jusqu'à nos jours. grossomodo, ce travail servira à améliorer l'espèce *dacryodes edulis*

## Problématique

L'amélioration d'une espèce que ce soit par le moyen des hybridations ou par des techniques simples de sélection, passe par la sauvegarde et ensuite par la connaissance de la diversité des populations de cette espèce. Nos espèces traditionnelles, généralement négligées, sont pour la plupart méconnues dans leur grande diversité. Pour cette raison, elles peuvent subir de manière imperceptible l'érosion génétique. C'est ainsi que certains écotypes ayant un haut potentiel disparaissent sans avoir été identifiés, privant de ce fait la sélection d'un matériel précieux. L'une des étapes dans la recherche sur l'amélioration d'une espèce végétale est la caractérisation qui a pour but la valorisation des différents écotypes de cette espèce. Le safoutier, arbre originaire des zones tropicales humides et sub- humides d'Afrique dont le fruit est le safou, d'une grande importance aussi bien économique qu'alimentaire, n'échappe pas à cette règle. C'est conscient de cette lacune que l'IRAD (Institut de Recherche Agricole pour le Développement) au Cameroun a procédé à une collection non ciblée d'accessions de safoutiers (ceci pour une conservation de la diversité des génotypes dans un endroit sécurisé) pour une caractérisation botanique de celles-ci ; ce qui permettra à long terme d'avoir de meilleures connaissances sur les différents écotypes.

## Revue de la littérature et quelques définitions

Le 2ème séminaire international sur la valorisation du safoutier et autres oléagineux non conventionnels effectué à l'ENSAI de Ngaoundéré rapporte que le développement et la promotion de la culture du safoutier se heurte à de nombreuses contraintes à savoir :

- l'insuffisance des connaissances scientifiques et techniques en ce qui concerne l'amélioration génétique, l'agronomie, la lutte contre les maladies, les technologies post-récolte ;
- le caractère très périssable du fruit, l'inexistence des technologies de conservation et de transformation, constituent des entraves majeures à la commercialisation ;
- l'absence d'informations sur le marché (quantités, prix), lacune aggravée par la non disponibilité de données statistiques fiables.

Les perspectives de développement de la filière safou, s'articulent en priorité autour de la recherche de solutions aux contraintes sus-mentionnées.

Plusieurs botanistes à diverses époques ont décrit l'espèce *Dacryodes edulis* (celle du safoutier) sous des noms scientifiques différents qui sont aujourd'hui tombés en synonymie. Comme exemple, on a :

- *Pachylobus edulis* G. Don (1982)
- *Canarium mubafo* (Ficalho) Engl. (1899)
- *Dacryodes edulis* (G. Don) H. J. Lam (1932)

### Différentes dénominations

Au Cameroun, les noms vernaculaires sont aussi variés que les dialectes et les tribus. Voici quelques uns :

Dschang : le-tsè, ékiep ; Bagangté : tchou ; Bafang : che ; Bamoun : youom ; Bassa-Ewondo, Eton : assa ; Bafia : kiyom ; Douala : sao ; Pygmée : senè. Au Congo et en République Démocratique du Congo : nsafou ; au Gabon atanga.

Safoutier est le nom français le plus original. Il dérive du nom vernaculaire Nsafou en lingala. Le terme prunier a définitivement cédé sa place au terme safoutier. En Anglais : Bush butter tree : en référence à sa richesse en acides gras et à son exploitation qui se faisait alors à l'état sauvage.

### Plan du travail

# MÉTHODES STATISTIQUES POUR LA CARACTÉRISATION

Diverses méthodes statistiques ont été développées dans la littérature pour la caractérisation. nous vous présenterons dans la suite quelques unes que nous avons jugées pertinentes pour atteindre les objectifs annoncés plus haut. Nous parlerons entre autres de L'Analyse en composantes principales, de l'analyse des correspondances multiples, de la segmentation. Au préalable nous présenterons quelques tests statistiques qui nous aideront dans nos analyses.

## 1.1 Tests statistiques

Dans cette partie nous explicitons le test du  $\chi^2$  pour la dépendance et celui de Shapiro pour la normalité.

subsection Test du  $\chi^2$

Soient  $X$  et  $Y$  2 variables statistiques ; notons  $\tilde{X}$  et  $\tilde{Y}$  les variables aléatoires associées respectivement à  $X$  et  $Y$ . Ces variables sont définies comme suit :

$$\tilde{X} : (\Omega, \mathcal{P}(\Omega), P) \longrightarrow (\mathcal{M}_X, \mathcal{P}(\mathcal{M}_X))$$

$$\tilde{Y} : (\Omega, \mathcal{P}(\Omega), P) \longrightarrow (\mathcal{M}_Y, \mathcal{P}(\mathcal{M}_Y))$$

où  $\Omega = \{1, \dots, n\}$  est l'ensemble des individus observés,  $(\Omega, \mathcal{P}(\Omega), P)$  l'espace probabilisé associé,  $P$  étant l'équiprobabilité ;  $\mathcal{M}_X$  (resp.  $\mathcal{M}_Y$ ) est l'ensemble des modalités de  $X$  (resp.  $Y$ ). Le test du  $\chi^2$  permet de s'assurer du caractère significatif de la liaison entre  $X$  et  $Y$ . Il est construit de la manière suivante :

l'hypothèse nulle est  $H_0$  :  $\tilde{X}$  et  $\tilde{Y}$  sont indépendantes en probabilités ;

l'hypothèse alternative est  $H_1$  : les variables  $\tilde{X}$  et  $\tilde{Y}$  ne sont pas indépendantes.

La statistique de test est alors :

$$\chi^2 = \sum_{l=1}^r \sum_{h=1}^c \frac{(n_{lh} - \frac{n_{l+}n_{+h}}{n})^2}{\frac{n_{l+}n_{+h}}{n}};$$

elle suit asymptotiquement (pour les grandes valeurs de  $n$ ), sous l'hypothèse  $H_0$ , une loi de  $\chi^2$  à  $(r-1)(c-1)$  degrés de liberté. On rejette donc  $H_0$  (et l'on conclut au caractère significatif de la liaison) si  $\chi^2$  dépasse une valeur particulière (valeur ayant une probabilité faible et fixée à priori ( en général 0,05 ) d'être dépassée par une loi de  $\chi^2$  à  $(r-1)(c-1)$  degrés de liberté).

### 1.1.1 Test de Shapiro-Wilk

Soit  $(x_1, \dots, x_n)$  un échantillon iid <sup>1</sup>, qu'on ordonne de façon croissante :

$$(x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}).$$

Si  $X$  suit une loi normale  $\mathcal{N}(m, \sigma^2)$ , on en déduit un échantillon  $(y_1, \dots, y_n)$  ordonné extrait de  $\mathcal{N}(0, 1)$ , avec  $y_i = \frac{x_i - m}{\sigma}$ .

**Définition 1.1.1.** On appelle scores normaux les quantités  $a_n(i) = E(X^{(i)})$ , où  $X^{(i)}$  est la  $i^{\text{ème}}$  observation ordonnée d'un échantillon de taille  $n$  extrait de  $\mathcal{N}(0, 1)$  :

$$X_{(1)} \leq \dots \leq X_{(i-1)} \leq X_{(i)} \leq X_{(i+1)} \leq \dots \leq X_{(n)}.$$

Les  $a_n(i)$  sont données dans la table des scores voir ([?], page 320).

L'idée du test de Shapiro-Wilk consiste à comparer deux statistiques  $T_1$  et  $T_2$  qui, sous l'hypothèse de normalité  $\mathcal{N}(m; \sigma)$  estiment toutes deux  $\sigma^2$ , et sous l'alternative, estiment des quantités différentes.

Shapiro et Wilk considèrent les statistiques de tests [?]

$$\begin{cases} T_1 = (n-1)S^2 \\ T_2 = \left( \sum_{i=1}^{n/2} a_n(i)(x_{(n-i+1)} - x_{(i)}) \right)^2, \end{cases} \quad (1.1)$$

où les  $a_n(i)$  sont les scores normaux, les  $x_{(i)}$  les observations ordonnées de façon croissante et  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  est la variance empirique de l'échantillon. La statistique de test est :

$$SW = \frac{T_2}{T_1}.$$

Des simulations ont montré que  $E(SW)$  était plus petit dans le cas non normal que le cas normal et  $V(SW)$  plus grand en non normal qu'en normal La région critique est :

$$W = \{SW < a\} \text{ telle que } P_{H_0}(W) = \alpha.$$

Le seuil  $a$  est lu dans la table de la loi normale donnée par [?],  $a$  est fonction de  $n$  et de  $\alpha$ .

## 1.2 Analyse en composantes principales

### 1.2.1 INTRODUCTION

Lorsqu'on étudie simultanément un nombre important de variables quantitatives (ne serait-ce que 4), comment en faire un graphique global ? La difficulté vient de ce que les individus étudiés ne sont plus représentés dans un plan, espace de dimension 2, mais dans un espace de

<sup>1</sup>indépendant et identiquement distribué

dimension plus importante (par exemple 4). L'objectif de l'Analyse en Composantes Principales (ACP) est de revenir à un espace de dimension réduite (par exemple 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales.

C'est la matrice des variances-covariances (ou celle des corrélations) qui va permettre de réaliser ce résumé pertinent, parce qu'on analyse essentiellement la dispersion des données considérées. De cette matrice, on va extraire, par un procédé mathématique adéquat, les facteurs que l'on recherche, en petit nombre. Ils vont permettre de réaliser les graphiques désirés dans cet espace de petite dimension (le nombre de facteurs retenus), en déformant le moins possible la configuration globale des individus selon l'ensemble des variables initiales (ainsi remplacées par les facteurs).

C'est l'interprétation de ces graphiques qui permettra de comprendre la structure des données analysées. Cette interprétation sera guidée par un certain nombre d'indicateurs numériques et graphiques, appelés aides à l'interprétation, qui sont là pour aider l'utilisateur à faire l'interprétation la plus juste et la plus objective possible.

L'analyse en Composantes Principales (ACP) est un grand classique de l'analyse des données pour l'étude exploratoire ou la compression d'un grand tableau  $n \times p$  de données quantitatives. L'ACP joue un rôle central car elle sert de fondement théorique aux autres méthodes de statistique multidimensionnelle dites *factorielles* qui en apparaissent comme des cas particuliers.

## 1.2.2 Représentation Vectorielle Des Données Quantitatives

### Notations

Soit  $p$  variables statistiques réelles  $X^j$  ( $j = 1 \dots p$ ) observées sur  $n$  individus  $i$  ( $i = 1 \dots n$ ) affectés des poids  $\omega_i$  :

$$\forall i = 1 \dots n : \omega_i > 0 \text{ et } \sum_{i=1}^n \omega_i = 1;$$

$$\forall i = 1 \dots n : x_i^j = X^j(i), \text{ mesure de } X^j \text{ sur le } i^{\text{ème}} \text{ individu}$$

Ces mesures sont regroupées dans une matrice  $X$  d'ordre  $(n \times p)$ .

	$X^1$	$\dots$	$X^j$	$\dots$	$X^p$
1	$x_1^1$	$\dots$	$x_1^j$	$\dots$	$x_1^p$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$i$	$x_i^1$	$\dots$	$x_i^j$	$\dots$	$x_i^p$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$n$	$x_n^1$	$\dots$	$x_n^j$	$\dots$	$x_n^p$

- A chaque individu  $i$  est associé le vecteur  $x_i$  contenant la  $i^{\text{ème}}$  ligne de  $X$  mise en colonne. C'est un élément d'un espace vectoriel noté  $E$  de dimension  $p$ ; nous choisissons  $\mathbb{R}^p$  muni de la base canonique  $\mathcal{E}$  et d'une métrique de matrice  $M$  lui conférant une structure d'espace euclidien :  $E$  est isomorphe à  $(\mathbb{R}^p, \mathcal{E}, M)$ ;  $E$  est alors appelé *espace des individus*.
- A chaque variable  $X^j$  est associé le vecteur  $x^j$  contenant la  $j^{\text{ème}}$  colonne centrée (la moyenne de la colonne est retranchée à toute la colonne) de  $X$ . C'est un élément d'un espace vectoriel noté  $F$  de dimension  $n$ ; nous choisissons  $\mathbb{R}^n$  muni de la base canonique  $\mathcal{F}$  et d'une métrique de matrice  $D$  diagonale des *poids* lui conférant une structure d'espace euclidien :  $F$  est isomorphe à  $(\mathbb{R}^n, \mathcal{F}, D)$  avec  $D = \text{diag}(\omega_1, \dots, \omega_n)$ ;  $F$  est alors appelé *espace des variables*.

### Interprétation Statistique De La Métrique Des Poids

L'utilisation de la métrique des poids dans l'espace des variables  $F$  donne un sens très particulier aux notions usuelles définies sur les espaces euclidiens. Ce paragraphe est la clé permettant de fournir les interprétations en termes statistiques des propriétés et résultats mathématiques.

$$\text{Moyenne empirique de } X^j : \bar{x}^j = \langle X e^j, 1_n \rangle_D = e^{j'} X' D 1_n = \sum_{i=1}^n \omega_i x_i^j.$$

$$\text{Barycentre des individus : } \bar{x} = X' D 1_n = (\bar{x}^1, \dots, \bar{x}^j, \dots, \bar{x}^p)'$$

$$\text{Matrice des données centrées : } \bar{X} = X - 1_n \bar{x}'.$$

$$\text{Ecart type de } X^j : \sigma_j = (x^{j'} D x^j)^{\frac{1}{2}} = \|x^j\|_D.$$

$$\text{Covariance de } X^j \text{ et } X^k : x^{j'} D x^k = \langle x^j, x^k \rangle_D.$$

$$\text{Matrice des covariances : } S = \sum_{i=1}^n \omega_i (x_i - \bar{x}) (x_i - \bar{x})' = \bar{X}' D \bar{X}.$$

$$\text{Corrélation de } X^j \text{ et } X^k : \frac{\langle x^j, x^k \rangle_D}{\|x^j\|_D \|x^k\|_D} = \cos \theta_D (x^j, x^k).$$

Par souci de simplicité des notations, on désigne toujours par  $x^j$  les colonnes de la matrice centrée  $\bar{X}$ . On considère donc que des vecteurs "variables" sont toujours centrés.

Ainsi, lorsque les variables sont centrées et représentées par des vecteurs de  $F$  :

- la *longueur* d'un vecteur représente un *écart-type*,
- le *cosinus* d'un angle entre deux vecteurs représente une *corrélation*.

## La Méthode

Les objectifs poursuivis par une ACP sont :

- La représentation graphique "optimale" des individus (lignes), minimisant les déformations du nuage des points, dans un sous-espace  $E_q$  de dimension  $q$  ( $q < p$ ),
- La représentation graphique des variables dans un sous-espace  $F_q$  en explicitant au "mieux" les liaisons initiales entre ces variables,
- La réduction de la dimension (compression), ou approximation de  $X$  par un tableau de rang  $q$  ( $q < p$ ).

Les derniers objectifs permettent d'utiliser l'ACP comme préalable à une autre technique préférant des variables orthogonales (régression linéaire) ou un nombre réduit d'entrées (réseaux de neurones).

Des arguments de type géométrique dans la littérature francophone, ou bien de type statistique avec hypothèses de normalité dans la littérature anglo-saxonne, justifient la définition de l'ACP. Nous adopterons ici une optique intermédiaire en se référant à un modèle "allégé" car ne nécessitant pas d'hypothèse "forte" sur la distribution des observations (normalité). Plus précisément, l'ACP admet des définitions équivalentes selon que l'on s'attache à la représentation des individus, à celle des variables ou encore à leur représentation simultanée.

### 1.2.3 Modèle

Les notations sont celles du paragraphe précédent :

- $X$  désigne le tableau des données issues de l'observation de  $p$  variables *quantitatives*  $X^j$  sur  $n$  individus  $i$  de *poids*  $\omega_i$ ,
- $E$  est l'espace des individus muni de la base canonique et de la métrique de matrice  $M$ ,
- $F$  est l'espace des variables muni de la base canonique et de la métrique des poids  $D = \text{diag}(\omega_1, \dots, \omega_n)$ .

De façon générale, un modèle s'écrit :

$$\text{Observation} = \text{Modèle} + \text{Bruit}$$

assorti de différents types d'hypothèses et de contraintes sur le modèle et sur le bruit.

En ACP, la matrice des données est supposée être issue de l'observation de  $n$  vecteurs aléatoires indépendants  $\{x_1, \dots, x_n\}$ , de même matrice de covariance, mais d'espérances différentes  $z_i$ , toutes contenues dans un sous-espace affine de dimension  $q$  ( $q < p$ ) de  $E$ . Dans ce modèle,  $E(x_i) = z_i$  est un paramètre spécifique attaché à chaque individu  $i$  et appelé *effet fixe*,

le modèle étant dit *fonctionnel*. Ceci s'écrit en résumé :

$$\left\{ \begin{array}{l} \{x_i; i = 1, \dots, n\}, n \text{ vecteurs aléatoires indépendants de } E, \\ x_i = z_i + \epsilon_i, i = 1, \dots, n \text{ avec } \begin{cases} E(\epsilon_i) = 0, \text{var}(\epsilon_i) = \sigma^2 \Gamma, \\ \sigma > 0 \text{ inconnu, } \Gamma \text{ régulière et connue,} \end{cases} \\ \exists A_q, \text{ sous-espace affine de dimension } q \text{ de } E \text{ tel que } \forall i, z_i \in A_q (q < p). \end{array} \right. \quad (1.2)$$

Soit  $\bar{z} = \sum_{i=1}^n \omega_i z_i$ . Les hypothèses du modèle entraînent que  $\bar{z}$  appartient à  $A_q$ . Soit donc  $E_q$  le sous-espace vectoriel de  $E$  de dimension  $q$  tel que :

$$A_q = \bar{z} + E_q.$$

les paramètres à estimer sont alors  $E_q$  et  $z_i, i = 1, \dots, n$ , éventuellement  $\sigma$ ;  $z_i$  est la part systématique, ou effet, supposée de rang  $q$ ; éliminer le bruit revient donc à réduire la dimension.

### estimation

**Proposition 1.2.1.** *L'estimation des paramètres de (??) est fournie par l'ACP de  $(X, M, D)$  c'est à dire par la décomposition en valeurs singulières de  $(\bar{X}, M, D)$  :*

$$\widehat{Z}_q = \sum_{k=1}^q \lambda_k^{1/2} u^k v^{k'} = U_q \Lambda^{1/2} V_q'.$$

- Les  $u_k$  sont les vecteurs propres  $D$ -orthonormés de la matrice  $\bar{X} M \bar{X}' D$  associés aux valeurs propres  $\lambda_k$  rangées par ordre décroissant.
- Les  $v_k$ , appelés vecteurs principaux, sont les vecteurs propres  $M$ -orthonormés de la matrice  $\bar{X}' D \bar{X} M = S M$  associés aux mêmes valeurs propres; ils engendrent des s.e.v. de dimension 1 appelés axes principaux.

Les estimations sont donc données par :

$$\begin{aligned} \widehat{\bar{z}} &= \bar{x}, \\ \widehat{Z}_q &= \sum_{k=1}^q \lambda_k^{1/2} u^k v^{k'} = U_q \Lambda^{1/2} V_q' = \bar{X} \widehat{P}_q', \\ \text{où } \widehat{P}_q &= V_q V_q' M \text{ est la matrice de projection } M\text{-orthogonale sur } \widehat{E}_q, \\ \widehat{E}_q &= \text{vect} v^1, \dots, v^q, \\ \widehat{E}_2 &\text{ est appelé plan principal,} \\ \widehat{z}_i &= \widehat{P}_q x_i + \bar{x}. \end{aligned}$$

### Remarques

i. Les espaces principaux sont uniques sauf, éventuellement, dans le cas de valeurs propres multiples.

ii. Si les variables ne sont pas homogènes (unités de mesure différentes, variances disparates), elles sont préalablement réduites :

$$\widetilde{X} = \bar{X} \Sigma^{-1/2} \text{ où } \Sigma = \text{diag} (\sigma_1^2, \dots, \sigma_p^2), \text{ avec } \sigma_j^2 = \text{Var} (X^j)$$

$\tilde{S}$  est alors la matrice  $R = \Sigma^{-1/2} S \Sigma^{-1/2}$  des corrélations.

Si on considère  $p$  variables statistiques centrées  $X^1, \dots, X^p$ , observées sur  $n$  individus de poids  $\omega_i$ , on pourra définir de manière équivalente à la précédente l'ACP de  $(X, M, D)$  comme la recherche des  $q$  combinaisons linéaires normées (standardisées) des  $X^j$ , non corrélées et dont la somme des variances soit maximale.

- Les vecteurs  $f^k = Mv^k$  sont les facteurs principaux. Ils permettent de définir les combinaisons linéaires des  $X^j$  optimales au sens ci-dessus.
- Les vecteurs  $c^k = \bar{X} f^k$  sont les composantes principales.
- Les variables  $C^k$  associées sont centrées, non corrélées et de variance  $\lambda_k$ ; ce sont les variables principales;

$$\begin{aligned} \text{cov}(C^k, C^l) &= (\bar{X} f^k)' D \bar{X} f^l = f^{k'} S f^l \\ &= v^{k'} M S M v^l = \lambda_l v^{k'} M v^l = \lambda_l \delta_k^l \end{aligned}$$

- Les  $f^k$  sont les vecteurs propres  $M^{-1}$ -orthonormés de la matrice  $MS$ .
- La matrice

$$C = \bar{X} F = \bar{X} M V = U \Lambda^{1/2}$$

est la matrice des composantes principales.

- Les axes définis par les vecteurs  $D$ -orthonormés  $u^k$  sont appelés axes factoriels.

## 1.2.4 Représentations graphiques

### Les variables

#### Projection

une variable  $X^j$  est représentée par la projection  $D$ -orthogonale de  $\widehat{Q}_q x^j$  sur le sous-espace  $F_q$  engendré par les  $q$  premiers axes factoriels. Les coordonnées de la projection  $D$ -orthogonale de  $x^j$  sur le sous-espace  $F_q$  sont les  $q$  premiers éléments de la  $j$ -ème ligne de la matrice  $V \Lambda^{1/2}$ .

#### Mesure de "qualité"

La qualité de la représentation de chaque  $x^j$  est donnée par le cosinus carré de l'angle qu'il forme avec sa projection :

$$[\cos \theta(x^j, \widehat{Q}_q x^j)]^2 = \frac{\|\widehat{Q}_q x^j\|_D^2}{\|x^j\|_D^2} = \frac{\sum_{k=1}^q \lambda_k (v_k^j)^2}{\sum_{k=1}^p \lambda_k (v_k^j)^2}.$$

#### Corrélations variable $\times$ facteurs

Ces indicateurs aident à l'interprétation des axes factoriels en exprimant les corrélations entre variables principales et initiales :

$$\text{cor}(X^j, C^k) = \cos \theta(x^j, c^k) = \cos \theta(x^j, u^k) = \frac{\langle x^j, u^k \rangle_D}{\|x^j\|_D} = \frac{\sqrt{\lambda_k} v_j^k}{\sigma_j}.$$

Ce sont les éléments de la matrice  $\Sigma^{-1/2} V \Lambda^{1/2}$ .

## Les individus

### Projection

Chaque individu  $i$  représenté par  $x_i$  est approché par sa projection  $M$ -orthogonale  $\widehat{z}_i^q$  sur le sous-espace  $\widehat{E}_q$  engendré par les  $q$  premiers vecteurs principaux  $\{v^1, \dots, v^q\}$ . ainsi, les coordonnées de la projection  $M$ -orthogonale de  $x_i - \bar{x}$  sur  $\widehat{E}_q$  sont les  $q$  premiers éléments de la  $i$ -ème ligne de la matrice  $C$  des composantes principales.

### Mesures de "qualité"

La "qualité globale" des représentations est mesurée par la part de dispersion expliquée :

$$r_q = \frac{\text{tr}SM\widehat{P}_q}{\text{tr}SM} = \frac{\sum_{k=1}^q \lambda_k}{\sum_{k=1}^p \lambda_k}.$$

La qualité de représentation de chaque  $x_i$  est donnée par le cosinus carré de l'angle qu'il forme avec sa projection :

$$[\cos \theta(x_i - \bar{x}, \widehat{z}_i^q)]^2 = \frac{\|\widehat{P}_q(x_i - \bar{x})\|_M^2}{\|x_i - \bar{x}\|_M^2} = \frac{\sum_{k=1}^q (c_i^k)^2}{\sum_{k=1}^p (c_i^k)^2}.$$

### Contributions

Les contributions de chaque individu à l'inertie de leur nuage :

$$\gamma_i = \frac{\omega_i \|x_i - \bar{x}\|_M^2}{\text{tr}SM} = \frac{\omega_i \sum_{k=1}^p (c_i^k)^2}{\sum_{k=1}^p \lambda_k}$$

Les contributions de chaque individu à la variance d'une variable principale :

$$\gamma_i^k = \frac{\omega_i (c_i^k)^2}{\lambda_k}.$$

Ces contributions permettent de déceler les informations les plus influentes.

### Individus supplémentaires

Ce sont les individus n'ayant pas participé aux calculs des axes principaux pour leur représentation, si on note  $s$  un tel vecteur (centré et éventuellement réduit), les coordonnées de cet individu dans la base des vecteurs principaux sont données par :

$$V_q' M(s - \bar{x}).$$

## 1.2.5 Choix de dimension

La qualité des estimations auxquelles conduit l'ACP dépend du choix de  $q$ , c'est à dire la dimension du sous-espace de représentation.

Nous avons rencontré plusieurs critères de choix pour  $q$  et nous exposerons quelques un dans la suite.

### Règle de Kaiser

D'après Kaiser, on ne conserve que les valeurs propres supérieures à la moyenne des valeurs propres. Dans le cas d'une ACP réduite, ne sont donc retenues que celles plus grandes que 1. Il faut cependant signaler que ce critère est implicitement utilisé par le logiciel SAS.

### Eboulis des valeurs propres

C'est le graphique présentant la décroissance des valeurs propres.

## 1.3 Analyse des Correspondances Multiples

Cette méthode est une généralisation de l'Analyse Factorielle des correspondances, permettant de décrire les relations entre  $p$  ( $p > 2$ ) variables qualitatives simultanément observées sur  $n$  individus. Elle est aussi souvent utilisée pour la construction de scores comme préalable à une méthode de classification nécessitant des données quantitatives.

### 1.3.1 Codages de variables qualitatives

#### Tableau disjonctif complet

Soit  $X$  une variable qualitative à  $c$  modalités. On appelle variable indicatrice de la  $k$ -ième modalité de  $x$  ( $k = 1, \dots, c$ ), la variable  $X_{(k)}$  définie par

$$X_{(k)}(i) = \begin{cases} 1 & \text{si } X(i) = \chi_k, \\ 0 & \text{sinon} \end{cases}$$

où  $i$  est un individu quelconque et  $\chi_k$  est la  $k$ -ième modalité de  $X$ . On notera  $n_k$  l'effectif de  $\chi_k$ .

On appelle matrice des indicatrices des modalités de  $X$ , et on notera  $\mathbf{X}$ , la matrice  $n \times c$  de terme général :

$$x_i^k = X_{(k)}(i).$$

On vérifie :

$$\sum_{k=1}^c x_i^k = 1, \forall i \text{ et } \sum_{i=1}^n x_i^k = n_k.$$

Considérons maintenant  $p$  variables qualitatives  $X^1, \dots, X^p$ . On note  $c_j$  le nombre de modalités de  $X^j$ ,  $c = \sum_{j=1}^p c_j$  et  $X_j$  la matrice des indicatrices de  $X^j$ .

On appelle alors tableau disjonctif complet la matrice  $\mathbf{X}$ ,  $n \times c$ , obtenue par concaténation des matrices  $X_j$  :

$$\mathbf{X} = [X_1 | \dots | X_p].$$

$\mathbf{X}$  vérifie :

$$\sum_{k=1}^c x_i^k = p, \forall i \text{ et } \sum_{i=1}^n \sum_{k=1}^c x_i^k = np.$$

D'autre part, la somme des éléments d'une colonne de  $\mathbf{X}$  est égale à l'effectif marginal de la modalité de la variable  $X^j$  correspondant à cette colonne.

### Tableau de Burt

On observe toujours  $p$  variables qualitatives sur un ensemble de  $n$  individus. On appelle tableau de Burt la matrice  $\mathbf{B}$ ,  $c \times c$ , définie par :

$$\mathbf{B} = \mathbf{X}'\mathbf{X}.$$

On peut écrire  $\mathbf{B} = [\mathbf{B}_{jl}] (j = 1, \dots, p; l = 1, \dots, p)$ ; chaque bloc  $\mathbf{B}_{jl}$ , de dimension  $c_j \times c_l$ , est défini par :

$$\mathbf{B}_{jl} = \mathbf{X}_j' \mathbf{X}_l.$$

Si  $j \neq l$ ,  $\mathbf{B}_{jl}$  est la table de contingence obtenue par croisement des variables  $X^j$  en lignes et  $X^l$  en colonnes. Si  $j = l$ , le bloc diagonal  $\mathbf{B}_{jj}$  est lui-même une matrice diagonale vérifiant :

$$\mathbf{B}_{jj} = \text{diag}(n_1^j, \dots, n_{c_j}^j).$$

La matrice  $\mathbf{B}$  est symétrique, d'effectifs marginaux  $n_i^j p$  et d'effectif total  $np^2$ .

## 1.3.2 Analyse Factorielle des Correspondances Multiples

### Définition

On considère  $p$  variables qualitatives ( $p \geq 3$ ) notées  $\{X^j; j = 1, \dots, p\}$ , possédant respectivement  $c_j$  modalités, avec  $c = \sum_{j=1}^p c_j$ . On suppose que ces variables sont observées sur les mêmes  $n$  individus, chacun affecté du poids  $1/n$ .

Soit  $\mathbf{X} = [\mathbf{X}_1 | \dots | \mathbf{X}_p]$  le tableau disjonctif complet des observations ( $\mathbf{X}$  est  $n \times c$ ) et  $\mathbf{B} = \mathbf{X}'\mathbf{X}$  le tableau de Burt correspondant ( $\mathbf{B}$  est carré d'ordre  $c$ , symétrique).

**Définition 1.3.1.** On appelle Analyse Factorielle des Correspondances Multiples (AFCM) des variables  $(X^1, \dots, X^p)$  relativement à l'échantillon considéré, l'AFC réalisée soit sur la matrice  $\mathbf{X}$  soit sur la matrice  $\mathbf{B}$ .

On note  $n_k^j (1 \leq j \leq p, 1 \leq k \leq c_j)$  l'effectif de la  $k$ -ième modalité de  $X^j$ ,  $D_j = \frac{1}{n} \text{diag}(n_1^j, \dots, n_{c_j}^j)$  et  $\Delta = \text{diag}(D_1 \dots D_p)$  ( $\Delta$  est carrée d'ordre  $c$  et diagonale).

**AFC du tableau disjonctif complet X**

Considérons les notations suivantes :

$$\begin{aligned} T &= X; \\ D_r &= \frac{1}{n} I_n; \\ D_c &= \frac{1}{p} \Delta; \\ A &= \frac{1}{p} X'; \\ B &= \frac{1}{p} X \Delta^{-1}. \end{aligned}$$

**ACP des profils-lignes**

**Proposition 1.3.1.** L'ACP des profils-lignes issue de l'AFC réalisée sur le tableau disjonctif complet de  $p$  variables qualitatives conduit à l'analyse spectrale de la matrice  $D_c^{-1}$ -symétrique et positive :

$$AB = \frac{1}{np} B \Delta^{-1}.$$

Il y a  $m$  ( $m \leq c - p$ ) valeurs propres notées  $\mu_k$ , ( $0 < \mu_k < 1$ ) rangées dans la matrice diagonale  $M$ .

La matrice des vecteurs propres  $D_c^{-1}$ -orthonormés associés se décompose en blocs de la façon suivante :

$$V = \begin{bmatrix} V_1 \\ \vdots \\ V_p \end{bmatrix};$$

chaque bloc  $V_j$  est de dimension  $c_j \times m$ .

La matrice des composantes principales s'écrit :

$$C_r = \sum_{j=1}^p X_j D_j^{-1} V_j.$$

La matrice des composantes principales permet de réaliser une représentation graphique des individus dans laquelle chacun apparaît, à un facteur près, comme le barycentre des  $p$  modalités qu'il a présentées.

ACP des profils-colonnes

**Proposition 1.3.2.** L'ACP des profils-colonnes issue de l'AFC réalisée sur le tableau disjonctif complet de  $p$  variables conduit à l'analyse spectrale de la matrice  $D_r^{-1}$ -symétrique et positive :

$$BA = \frac{1}{np} X \Delta^{-1} X' = \frac{1}{np} \sum_{j=1}^p X_j D_j^{-1} X_j'.$$

La matrice des vecteurs propres  $D_r^{-1}$ -orthonormés vérifie :

$$U = B V M^{-1/2}.$$

La matrice des composantes principales s'écrit :

$$C_c = p\Delta^{-1}VM^{-1/2};$$

elle se décompose en blocs sous la forme :

$$C_c = \begin{bmatrix} C_1 \\ \vdots \\ C_p \end{bmatrix}.$$

Chaque bloc  $C_j$ , de dimension  $c_j \times m$ , fournit en lignes les coordonnées des modalités de la variable  $X^j$  permettant la représentation graphique simultanée.

### AFC du tableau de Burt $\mathcal{B}$

Le tableau de Burt  $\mathcal{B} = X'X$ , carré d'ordre  $c$  étant symétrique, les profils-lignes et les profils-colonnes sont identiques ; on ne considère donc ici qu'une seule ACP.

En utilisant le tilde dans ce cas, les matrices usuelles de l'AFC deviennent :

$$\begin{aligned} \tilde{T} &= \mathcal{B}; \\ \tilde{D}_r &= \tilde{D}_c = \frac{1}{p}\Delta = D_c; \\ \tilde{A} &= \tilde{B} = \frac{1}{np}\mathcal{B}\Delta^{-1} = AB. \end{aligned}$$

**Proposition 1.3.3.** L'ACP des profils-lignes (ou des profils colonnes) issue de l'AFC réalisée sur le tableau de Burt associé à  $p$  variables qualitatives conduit à l'analyse spectrale de la matrice  $\tilde{D}_c^{-1}$ -symétrique et positive :

$$\tilde{A}\tilde{B} = [AB]^2.$$

Elle admet pour matrice de vecteurs propres  $\tilde{D}_c^{-1}$ -orthonormés  $\tilde{U} = \tilde{V} = V$ .

Les valeurs propres associées vérifient  $\nu_k = \mu_k^2$ .

La matrice des composantes principales s'écrit :

$$\tilde{C}_r = \tilde{C}_c = C_c M^{1/2} = \begin{bmatrix} C_1 \\ \vdots \\ C_p \end{bmatrix} M^{1/2}.$$

La matrice  $\tilde{C}_r$  fournit les coordonnées permettant la représentation simultanée des modalités de toutes les variables (on ne peut pas faire de représentation des individus si l'on fait l'AFC du tableau de Burt).

# ETUDE DESCRIPTIVE DES DONNÉES SUR LE SAFOU

---

## 2.1 Présentation des données

### Origine des données et présentation des variables

*Les accessions de safoutiers ont été collectées dans la ville de Brazzaville au Congo et dans les provinces de l'Ouest, du Sud-Ouest, du Centre, du Sud, du Littoral au Cameroun. Après la collecte elles ont été conservées à la banque des gènes de Barombi-Kang, une station de l'Institut de Recherche Agricole pour le Développement (IRAD) située dans la province du Sud-Ouest au Cameroun. La collecte était non ciblée<sup>1</sup> ceci pour une préservation de la diversité.*

*Les variables mesurées sont décrites ci-dessous :*

#### 1. Variables Qualitatives

- *Provenance* : variable à 6 modalités représentant les différents sites où on a collecté les accessions de safoutier ; il s'agit de la ville de Brazzaville au Congo et des provinces de l'Ouest, du Sud-Ouest, du Centre, du Sud, du Littoral.
- *Sexe* : variable à 2 modalités indiquant le sexe de l'Accession ; M = Mâle et F = Femelle.
- *Mode de ramification* : variable à 3 modalités indiquant le mode de ramification des accessions ; D = dressé, H = horizontal et R = retombant.
- *Aspect écorce* : variable à 2 modalités indiquant l'aspect de l'écorce des accessions ; L = Lisse et R = Rugueux.
- *Aspect limbe* : variable à 2 modalités indiquant l'aspect du limbe ; G = gaufré et L = lisse
- *Couleur* : variable à 2 modalités indiquant la couleur des feuilles C = vert-clair et S = vert-sombre

#### 2. Variables Quantitatives

- *long inflore* : elle représente la longueur de l'inflorescence.
- *Nomb réitérations* : elle représente le nombre de ramifications de l'inflorescence.
- *Nomb fleurs* : elle représente le nombre de fleurs.
- *Nfleurs/inflo* : elle représente le nombre de fleurs par inflorescence.
- *Long feuille* : elle représente la longueur des feuilles.

---

<sup>1</sup>lorsqu'elle est ciblée, on collecte en fonction des caractéristiques requises

- Nomb foliole : elle représente le nombre de foliole.
- Long foliole : elle représente la longueur des folioles.
- Larg foliole : elle représente la largeur des folioles.

### Variables quantitatives

#### – Analyses Univariées :

Les statistiques élémentaires sont contenues dans le tableau ci-dessous.

variable	min	moyenne	variance	max	% de NA
Long.inflo	9	18.48	45.60477	45	2,91
Nomb.reiterations	0	1.112	8.67	45	3.18
Nomb.fleurs	9	70.61	1271.012	240	2.91
Nfleurs.inflo	6	61.48	1029.209	197	3.44
Long.feuille	27.40	48.67	36.98	65.40	0
Nomb.foliole	9	15.89	2.25	21	0.26
Long.foliole	10.50	17.18	2.71	23.40	0
Larg.foliole	3.50	6.017	11.008	52	0.26

TAB. 2.1 – Statistiques élémentaires des variables quantitatives

On remarque là une grande hétérogénéité entre les 8 variables considérées : ordre de grandeur distinct pour les moyennes, les variances, les minima et les maxima.

#### – Analyse Bivariées

Nous avons utilisé la matrice des corrélations (??) en ayant préalablement imputé les valeurs manquantes.

	Long .inflo	Nomb. Reiterations	Nomb. fleurs	Nfleurs .inflo	Long. feuille	Nomb. foliole	Long. foliole	Larg. foliole
Long.inflo	1	0.34	<b>0.71</b>	<b>0.71</b>	0.09	0.07	0.10	-0.04
Nomb.reiterations		1	0.31	0.35	0.07	0.02	0.07	-0.03
Nomb.fleurs			1	<b>0.98</b>	0.11	0.09	0.10	-0.02
Nfleurs.inflo				1	0.12	0.11	0.10	0.01
Long.feuille					1	0.51	0.61	0.12
Nomb.foliole						1	0.34	0.05
Long.foliole							1	0.15
Larg.foliole								1

TAB. 2.2 – Matrice des corrélations entre les variables

Il ressort de ce tableau (??) que toutes les corrélations linéaires à l'exception de certaines avec la variable Larg.foliole sont positives (ceci signifie que toutes les autres variables

varient, en moyenne, dans le même sens), la plupart étant très faibles (0.15, ..., 0.01) et quelques unes par contre très fortes (0.71, ..., 0.98).

### Variables Qualitatives

#### – Analyse Univariée

Les résumés des variables qualitatives sont contenus dans le tableau (??) ci-dessous :

	Fréquence par modalités	% de NA
Provenance	BR :3.71 ; CE : <b>40.84</b> ;LT :15.38 OU :20.68 ; SU : <b>1.32</b> ;SW :18.03	0
Sexe	F :47,21 ; M :50.13	2.65
Mode.de.Ramification	D : <b>71.35</b> ; H :27.85 ; R : <b>0.79</b>	0
Aspect.ecorce	L :68.43 ; R :31.56	0
Aspect.limbe	G :42.7 ; L :57.29	0
Couleur	C :47.48 ; S :52.51	0

TAB. 2.3 – Résumer des variables qualitatives

Ce tableau (??) nous révèle que la prospection d'accessions de safoutiers était intense dans la province du centre et très faible dans la province du sud ; au niveau du mode de ramification, les Dressés se sont taillé la grande part en laissant juste une petite part aux Retombants.

#### – Analyse Bivariée

On effectuera ici le test du KHI-2 pour la signficativité des éventuelles liaisons entre les variables et des diagrammes en barres croisées pour voir leurs répartitions. Les p-value de ce test sont contenues dans le tableau (??) suivant :

	Sexe	Mode.de ramification	Aspect.ecorce	Aspect.limbe	Couleur
Provenance	0.94	0.508	0.11	0.27	0.06
Sexe		0.67	0.53	0.97	0.54
Mode.de.ramification			0.77	0.69	0.79
Aspect.ecorce				0.34	0.57
Aspect.limbe					< 2.2 * 10 <sup>-16</sup>

TAB. 2.4 – p-value du test du KHI-2 entre les variables

On constate donc d'après (??) que la liaison entre les variables Aspect limbe et Couleur est hautement significative contrairement aux autres variables qui sont indépendantes. Les diagrammes en barres croisées de certaines variables sont données ci-dessous ;

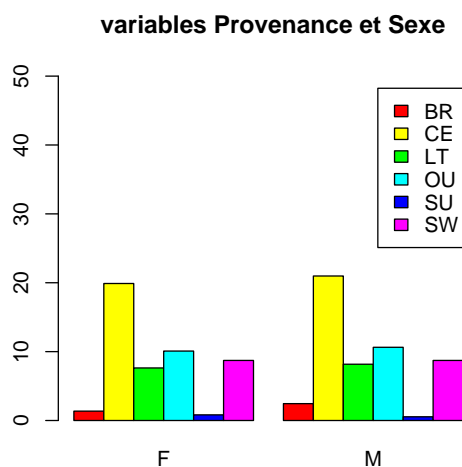


FIG. 2.1 – diagramme en barres croisées des variables Sexe et Provenance

La figure (??) représente la répartition des sexes par provenance et on constate que la distribution des sexes par provenance est sensiblement la même.

De même, la figure (??) nous révèle que les accessions ayant un aspect du limbe lisse (L) ont tendance à avoir une couleur de feuille vert-clair (C) et parallèlement celles ayant un aspect du limbe gaufré (G) ont tendance à avoir une couleur de feuille vert-sombre (S). Ceci conforte les résultats du test du KHI-2 précédent relativement à ces deux variables.

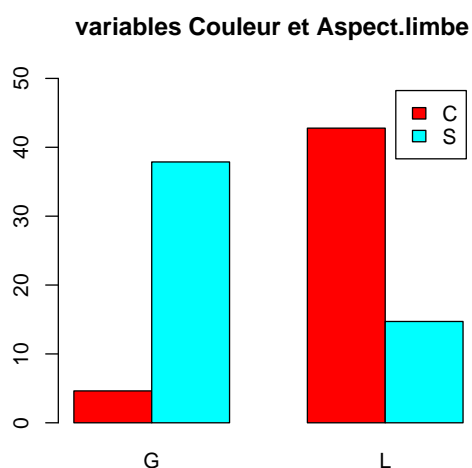


FIG. 2.2 – diagramme en barres croisées des variables Aspect.limbe et Couleur

# APPLICATIONS DES MÉTHODES À NOTRE JEU DE DONNÉES ET RÉSULTATS OBTENUS

## 3.1 Application de l'ACP au jeu de données quantitatives

Les valeurs propres du tableau ci-dessous (??) sont celles de la matrice des corrélations (??)

Facteur	Val.pr	Pct.Var	Pct.Cum
f1	2.900	36	36
f2	1.909	24	60
f3	0.975	12	72
f4	0.824	10	82
f5	0.653	8	90
f6	0.366	5	95
f7	0.350	4	99
f8	0.023	1	100

TAB. 3.1 – tableau des valeurs propres de la matrice des corrélations

### Interprétation

Chaque ligne du tableau ?? correspond à une variable virtuelle (les facteurs) dont la colonne Val.Pr (valeur propre) fournit la variance (en fait, chaque valeur propre représente la variance du facteur correspondant). La colonne Pct.Var, ou pourcentage de variance, correspond au pourcentage de variance de chaque ligne par rapport au total. La colonne **Pct.Cum** représente le cumul de ces pourcentages.

On remarque que le nuage de points en dimension 8 reste le même et sa dispersion globale n'a pas changé car la somme des valeurs propres ci-dessus est égale à la somme des variances des variables initiales. Il s'agit d'un simple changement de base dans un espace vectoriel.

C'est la répartition de cette dispersion, selon les nouvelles variables que sont les facteurs, ou composantes principales, qui se trouve modifiée.

Le critère de Kaiser nous permet de ne retenir que les 2 premiers facteurs au vu des valeurs propres car seules les 2 premières valeurs propres sont  $> 1$ . De même, l'ébouli des valeurs propres (??) nous suggère la même dimension.

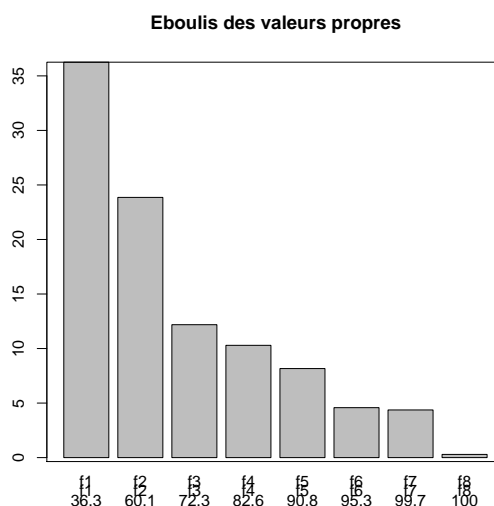


FIG. 3.1 – Ebouli des valeurs propres

Par conséquent, les graphiques en dimension 2 qui seront présentés ci-dessous résument presque parfaitement la configuration réelle des données se trouvant en dimension 8 : l'objectif (résumé pertinent des données en petite dimension) est donc atteint.

### 3.1.1 Résultats sur les variables

Le résultat fondamental concernant les variables est le tableau des corrélations variables-facteurs (??). Il s'agit des coefficients de corrélation linéaire entre les variables initiales et les facteurs. Ce sont ces corrélations qui permettront de donner un sens aux facteurs (de les interpréter). nous ne nous intéresserons qu'aux 2 premiers facteurs.

Les deux colonnes du tableau (??) permettent, tout d'abord, de réaliser le graphique des variables (??).

Mais ces deux colonnes permettent également de donner une signification aux facteurs (donc aux axes des graphiques). Comme complément à ce tableau on a le tableau des contributions des variables aux facteurs :

#### Interprétation

Ainsi, on voit que le premier facteur est corrélé positivement avec chacune des 8 variables initiales. Les variables ayant une importance dans la définition de l'axe 1 sont au vu des corrélations : "Nfleurs.inflo", "Nomb.fleurs" et dans une certaine mesure "Long.inflore". Ceci s'illustre

	f1	f2
Long.inflore	0.481	-0.168
Nomb.reiterations	0.292	-0.098
Nomb.fleurs	0.535	-0.182
Nfleurs.inflo	0.540	-0.176
Long.feuille	0.204	0.584
Nomb.foliole	0.166	0.479
Long.foliole	0.189	0.535
Larg.foliole	0.011	0.204

TAB. 3.2 – Corrélation Variables - Facteurs

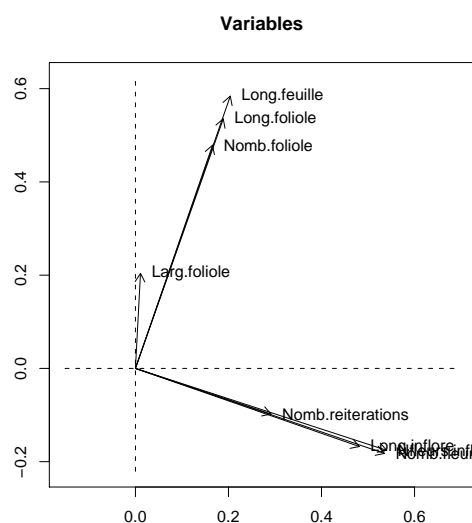


FIG. 3.2 – représentation des variables dans le plan principal

	f1	f2	f>2
Long.inflore	232	28	86
Nomb.reiteration	86	9	230
Nomb.fleurs	287	33	33
Nfleurs.inflo	292	31	30
Long.feuille	41	341	72
Nomb.foliole	27	231	151
Long.foliole	35	286	110
Larg.foliole	0	41	289

TAB. 3.3 – tableau des cotributions des variables aux facteurs

bien avec le tableau des contributions des variables aux facteurs. Ces trois variables renvoient à l'inflorescence. En ce qui concerne l'axe 2, il oppose d'une part, les variables "Long.inflore", "Nomb.reiteration", "Nomb.fleurs", "Nfleurs.info" (corrélations négatives) et d'autre part, les variables "Long.feuille", "Nomb.foliole" (corrélations positives). Il s'agit donc d'un axe d'opposition entre la fleur et la feuille. Cependant, ce n'est qu'une illusion due à la projection des variables dans ce sous espace car d'après le tableau (??), les variables "Long.feuille", "Nomb.foliole", "Long.foliole" contribuent le plus à la création de l'axe2 donc sont par conséquent les plus importantes pour son interprétation contrairement aux autres qui y contribuent très faiblement. On peut donc conclure que l'axe2 renvoie à la feuille.

Cette interprétation pourra être précisée avec les graphiques et tableaux relatifs aux individus que nous présenterons dans la partie suivante.

### 3.1.2 Résultats sur les individus

La taille des individus de notre fichier de données étant grande, nous donnerons dans cette section juste un résumé des sorties R relativement aux individus dans l'acp réalisée. le tableau récapitulatif suivant donne les divers résultats pour les 20 premiers individus de notre tableau de données quantitatives. Néanmoins les conclusions auxquelles nous aboutirons seront faites à l'aide des résultats sur tous les individus.

	axe1	axe2	cont1	cont2	cosca1	cosca2
1	0,434	-2,652	0	10	19	716
2	2,251	-0,169	5	0	443	445
3	-1,815	-0,165	3	0	872	879
4	-2,281	-3,068	5	13	209	588
5	4,762	1,511	21	3	898	988
6	-0,126	-0,04	0	0	16	18
7	2,2	-1,232	4	2	507	666
8	-0,197	-2,917	0	12	4	829
9	0,067	-0,046	0	0	1	2
10	-0,483	2,008	0	6	52	942
11	3,241	1,674	10	4	418	529
12	-0,487	-1,703	0	4	31	406
13	-0,676	0,669	0	1	130	257
14	-2,175	-0,174	4	0	833	839
15	3,478	-0,494	11	0	407	415
16	1,095	0,36	1	0	178	197
17	1,474	-0,316	2	0	396	414
18	-0,199	0,327	0	0	6	23
19	-1,296	1,403	2	3	223	484
20	1,303	-0,387	2	0	351	382

TAB. 3.4 – tableau récapitulatif de l'acp sur les individus

### Interprétation

la première colonne du tableau (??) représente les individus qui ici sont les accessions de safoutiers. Chaque individu représente 1 élément sur 377, d'où un poids<sup>1</sup> de  $1/377 = 0.0026$ .

Les 2 colonnes suivantes (axe1,axe2) fournissent les coordonnées des individus sur les deux premiers axes (facteurs) et permettent de réaliser le **graphique des individus** (??).

Les 2 colonnes suivantes (cont1,cont2) fournissent les **contributions** des individus à diverses dispersions : cont1 donne les contributions des individus à la variance selon l'axe1 et cont2 celles des individus à la variance selon l'axe2. elles sont exprimées en millièmes (chaque colonne somme à 1000) et permettent de repérer les individus les plus importants au niveau de chaque axe.

Les 2 dernières colonnes sont des cosinus carrés fournissant la qualité de représentation de chaque individu sur chaque dimension. ainsi, cosca1 est celle de l'axe1 et cosca2 celle du plan principal (pour avoir la qualité de représentation sur l'axe 2 il suffit de faire la différence  $\text{cosca2} - \text{cosca1}$ ). La méthodologie qu'on adoptera pour l'interprétation proprement dite est la suivante :

- On repère les individus "bien représentés" dans le plan en utilisant la convention  $\text{cosca2} \geq 900$
- On repère leurs contributions sur les axes pour déterminer les plus importants sur chaque axe .
- Peaufiner l'interprétation de la section précédente.

La représentation graphique des individus est donnée par la figure(??)

en regardant cette figure les individus bien représentés dans le plan et particulièrement sur l'un des axes sont visibles. On s'intéressera juste aux individus extrêmes des axes.

le graphique récapitulatif de l'ACP est donné par (??)

## 3.2 Application de l'ACM au jeu de données qualitatives

La figure suivante illustrera les relations entre les variables du jeu de données qualitatives. elle s'obtient en effectuant l'AFC du tableau de Burt associé au jeu de données.

---

<sup>1</sup>une pondération

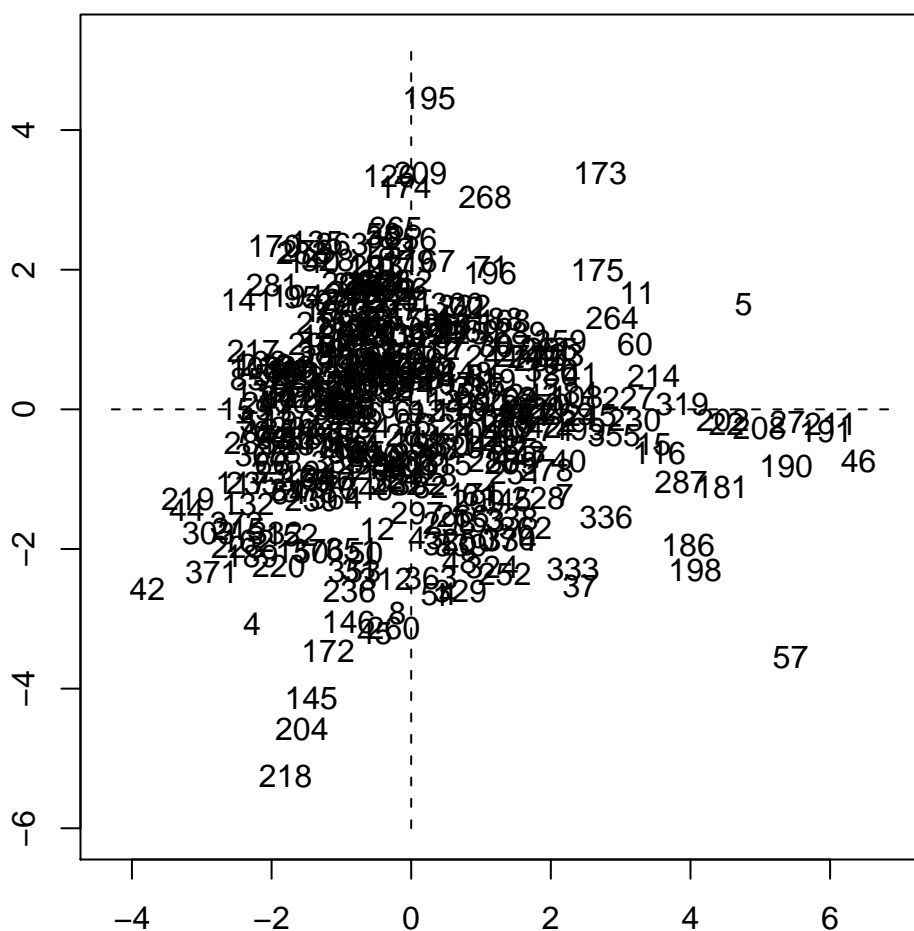


FIG. 3.3 – représentation des individus dans le plan principal

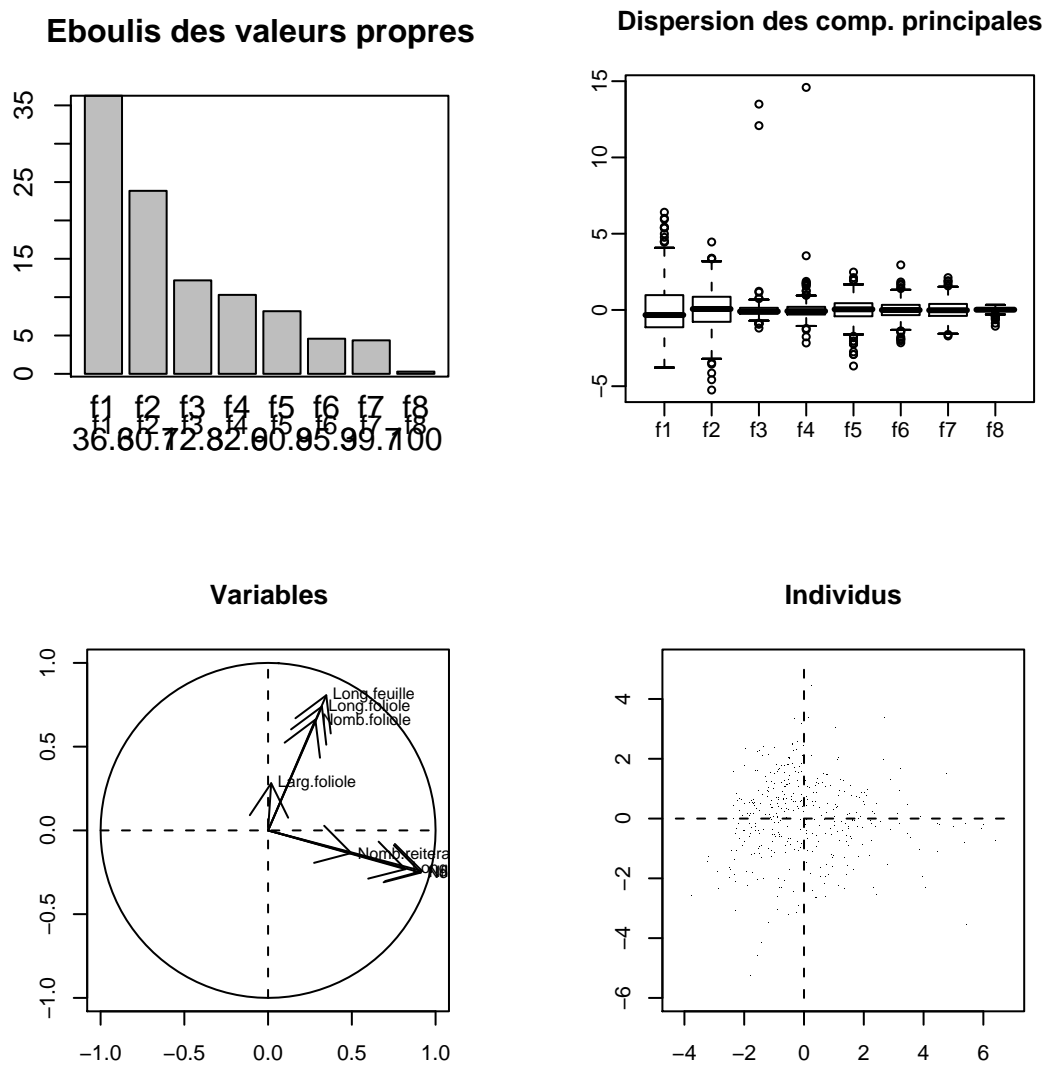


FIG. 3.4 – Récapitulatif de l'ACP

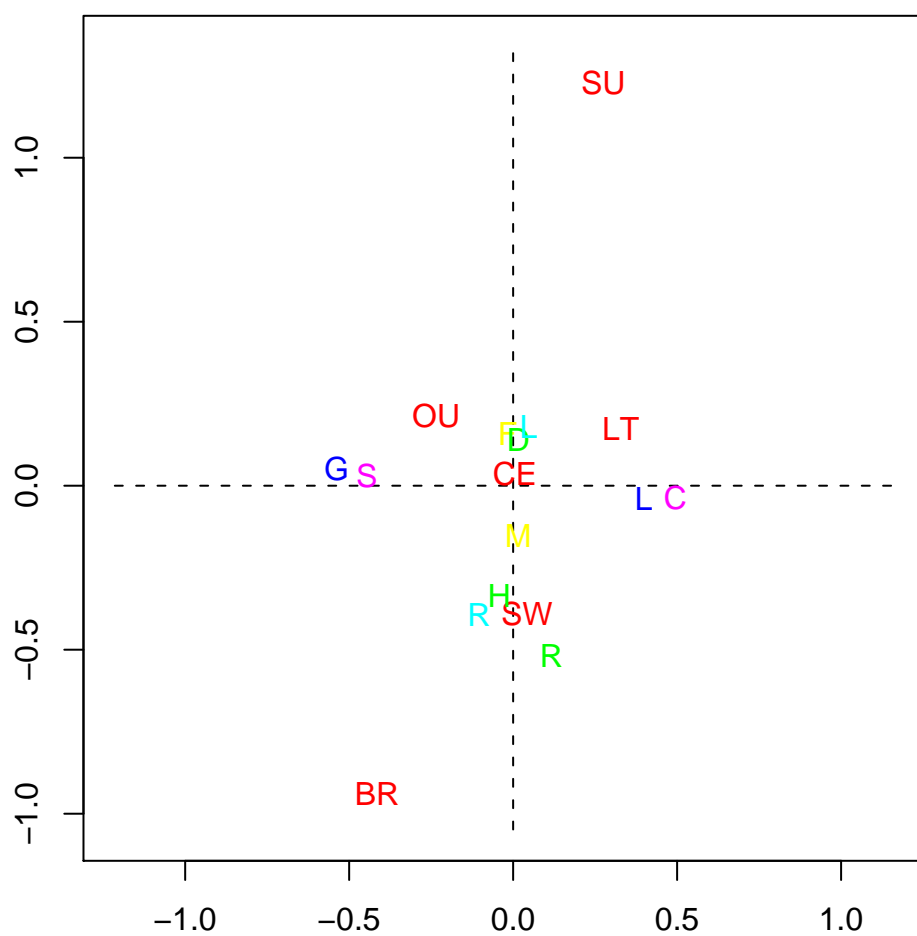


FIG. 3.5 – AFCM du tableau de Burt

# ANNEXE

---

## 4.1 Outils Algébriques

*Cette section se propose de rassembler des notations et rappels d'algèbre linéaire ainsi que quelques compléments mathématiques du niveau du premier cycle des Universités.*

*Dans tout ce qui suit,  $E$  et  $F$  sont deux espaces vectoriels réels munis respectivement des bases canoniques  $\mathcal{E} = \{e_j; j = 1, \dots, p\}$  et  $\mathcal{F} = \{f_i; i = 1, \dots, n\}$ . On note indifféremment soit un vecteur de  $E$  ou de  $F$ , un endomorphisme de  $E$ , ou une application linéaire de  $E$  dans  $F$ , soit leurs représentations matricielles dans les bases définies ci-dessus.*

### 4.1.1 Matrices

#### Notations

*La matrice d'ordre  $(n \times p)$  associée à une application linéaire de  $E$  vers  $F$  est décrite par un tableau :*

$$A = \begin{bmatrix} a_1^1 & \cdots & a_1^j & \cdots & a_1^p \\ \vdots & & \vdots & & \vdots \\ a_i^1 & \cdots & a_i^j & \cdots & a_i^p \\ \vdots & & \vdots & & \vdots \\ a_n^1 & \cdots & a_n^j & \cdots & a_n^p \end{bmatrix}.$$

*On note par la suite*

$$a_i^j = [A]_i^j \text{ le terme général de la matrice,}$$

$$a_i = [a_i^1, \dots, a_i^p]' \text{ un vecteur-ligne mis en colonne,}$$

$$a^j = [a_1^j, \dots, a_n^j]' \text{ un vecteur colonne.}$$

#### Types de matrices

*Une matrice est dite :*

- vecteur-ligne (colonne) si  $n = 1$  ( $p = 1$ ),
- vecteur-unité d'ordre  $p$  si elle vaut  $1_p = [1, \dots, 1]'$ ,

- scalaire si  $n = 1$  et  $p = 1$ ,
- carrée si  $n = p$ .

Une matrice carrée est dite :

- identité ( $I_p$ ) si

$$a_i^j = \delta_i^j = \begin{cases} 0 & \text{si } i \neq j \\ 1 & \text{si } i = j \end{cases}$$

- diagonale si  $a_i^j = 0$  lorsque  $i \neq j$ ,
- symétrique si  $a_i^j = a_j^i, \forall (i, j)$ ,
- triangulaire inférieure (supérieure) si  $a_i^j = 0$  lorsque  $i < j$  ( $i > j$ ).

### Matrice partitionnée en blocs

Matrices dont les éléments sont eux-mêmes des matrices. Exemple :

$$A (n \times p) = \begin{bmatrix} A_1^1 (r \times s) & A_1^2 (r \times (p-s)) \\ A_2^1 ((n-r) \times s) & A_2^2 ((n-r) \times (p-s)) \end{bmatrix}.$$

## Opérations sur les matrices

### Somme

$$[A + B]_i^j = a_i^j + b_i^j$$

pour  $A$  et  $B$  de même ordre ( $n \times p$ ).

### Multiplication par un scalaire

$$[\alpha A]_i^j = \alpha a_i^j$$

pour  $\alpha \in \mathbb{R}$ .

### Transposition

$$[A']_i^j = a_j^i$$

,  $A'$  est d'ordre ( $p \times n$ ).

$$(A')' = A; (A + B)' = A' + B'; (AB)' = B' A'; \begin{bmatrix} A_1^1 & A_1^2 \\ A_2^1 & A_2^2 \end{bmatrix}' = \begin{bmatrix} A_1^{1'} & A_2^{1'} \\ A_1^{2'} & A_2^{2'} \end{bmatrix}.$$

### Produit

$$[AB]_i^j = a_i^k b^k_j \text{ avec } A_{(n \times p)}, B_{(p \times q)}, \text{ et } AB_{(n \times q)},$$

et pour des matrices par blocs :

$$\begin{bmatrix} A_1^1 & A_1^2 \\ A_2^1 & A_2^2 \end{bmatrix} \begin{bmatrix} B_1^1 & B_1^2 \\ B_2^1 & B_2^2 \end{bmatrix} = \begin{bmatrix} A_1^1 B_1^1 + A_1^2 B_2^1 & A_1^1 B_1^2 + A_1^2 B_2^2 \\ A_2^1 B_1^1 + A_2^2 B_2^1 & A_2^1 B_1^2 + A_2^2 B_2^2 \end{bmatrix}$$

sous réserve de compatibilité des dimensions.

## Propriétés des matrices carrées

la trace et le déterminant sont des notions intrinsèques, qui ne dépendent pas des bases de représentation choisies, mais uniquement de l'application linéaire sous-jacente.

### Trace

Par définition, si  $A$  est une matrice  $(p \times p)$ ,

$$\text{tr}A = \sum_{j=1}^p a_j^j,$$

### **Déterminant**

On note  $|A|$  le déterminant de la matrice carrée  $A(p \times p)$ . ce dernier vérifie lorsque la matrice  $A$  est triangulaire ou diagonale la relation :

$$|A| = \prod_{j=1}^p a_j^j.$$

**quelques définitions**

Les applications statistiques envisagées dans ce cours ne s'intéressent qu'à des types particuliers de matrices.

**Théorème 4.1.1.** Une matrice  $A$  réelle symétrique admet  $p$  valeurs propres réelles. Ses vecteurs propres peuvent être choisis pour constituer une base orthonormée de  $E$ ;  $A$  se décompose en :

$$A = V\Lambda V' = \sum_{k=1}^p \lambda_k v^k v^{k'}$$

Où  $V$  est une matrice orthogonale  $[v^1, \dots, v^p]$  des vecteurs propres orthonormés associés aux valeurs propres  $\lambda_k$ , rangées par ordre décroissant dans la matrice diagonale  $\Lambda$ .

**Théorème 4.1.2.** Une matrice  $A$  réelle  $M$ -symétrique admet  $p$  valeurs propres réelles. Ses vecteurs propres peuvent être choisis pour constituer une base  $M$ -orthonormée de  $E$ ;  $A$  se décompose en :

$$A = V\Lambda V' M = \sum_{k=1}^p \lambda_k v^k v^{k'} M$$

où  $V = [v^1, \dots, v^p]$  est une matrice  $M$ -orthogonale ( $V' M V = I_p$  et  $V V' = M^{-1}$ ) des vecteurs propres associés aux valeurs propres  $\lambda_k$ , rangées par ordre décroissant dans la matrice diagonale  $\Lambda$ .

Les décompositions ne sont pas uniques : pour une valeur propre simple (de multiplicité 1) le vecteur propre normé est défini à un signe près, tandis que pour une valeur propre multiple, une infinité de bases  $M$ -orthonormées peuvent être extraites du sous-espace propre unique associé.

Le rang de  $A$  est aussi le rang de la matrice  $\Lambda$  associée et donc le nombre (répétées avec leurs multiplicités) de valeurs propres non nulles.

Par définition, si  $A$  est positive, on note la racine carrée de  $A$  :

$$A^{1/2} = \sum_{k=1}^p \sqrt{\lambda_k} v^k v^{k'} M = V \Lambda^{1/2} V' M.$$

## Propriété 4.1.1. 4.2 Codes R utilisés

dans cette section, nous avons utilisé dans la mesure du possible les fonctions prédéfinies de R et au cas échéant nous avons implémenté les nôtres et aussi celles de

```
# Importation du tableau " données safou "
# Après s'être placé dans le répertoire contenant
# le fichier à l'aide de la fonction
```

```
setwd("chemin d'accès au fichier")
```

```
# Importation proprement dite :

library( xlsReadWrite)

Tab = read.xls( "données safou")

# Analyse Descriptive

#Extraction des variables quantitatives

Tab1 = tab [, sapply( tab,is.numeric )]

# Extraction des variables qualitatives

Tab2=tab [, sapply(tab,is.factor)]

# Résumer des variables quantitatives

apply( Tab1,2,summary)

#Résumer des variables qualitatives

apply( Tab2,2, table)

# Pour réaliser les diagrammes croisés en boites parallèles
#des variables qualitatives,on utilise la fonction suivante qu'on
#appliquera aux variables du tableau Tab2

#elle fournit aussi la table de contingence

ana.bi.quali= fonction(var1,var2, table=Tab2, ident=F)
{
vec1<-table[,var1]
vec2<-table[,var2]
nlev1<-length(attributes(vec1)$levels)
nlev2<-length(attributes(vec2)$levels)
conting<-matrix(0,nlev1,nlev2)
cont<-by(vec1,vec2,summary)
for (i in 1 :nlev2)
{
conting[,i]<-cont[[i]]
}
nomi<-attributes(vec1)$levels
nomj<-attributes(vec2)$levels
dimnames(conting)<-list(nomi,nomj)
```

```

print("_____",quote = FALSE)
print(paste("Table de contingence",var1,"x",var2,"(en effectif)",quote = FALSE)
print(conting)
print("_____",quote = FALSE)
contingp<-contingp<-round(100*conting/sum(conting),2)
print("_____",quote
= FALSE)
print(paste("Table de contingence",var1,"x",var2,"(en pourcentage du total)",quote =
FALSE)
print(contingp)
print("_____",quote
= FALSE)
x11 ( )
barplot (contingp, beside = TRUE,col = rainbow(nlev1),legend = rownames(contingp),
ylim = c(0, max(50,max(contingp)+5)))
title(paste("Diagramme en barres croisées des variables",var1,"et",var2,"\n(en pourcen-
tage)"))
}
# Analyse Multi –dimensionnelles
# fonction réalisant le tableau de Burt de deux tableaux de variables qualitatives
BurtR= fonction(x, y, nomidx = "paste", nomidy = "paste", indx = F, indy = F)
{
#_____
# Generation d'un tableau de Burt ou d'une partie de tableau de Burt
# a partir des tableaux qualitatifs x et y
# Si y est manquant, alors tout se passe comme si x=y et le tableau
# construit est un tableau de Burt symetrique.
# Si les noms des modalites ne sont pas fournis,
# ils sont generes automatiquement.
# Si x et y sont des tableaux d'indicateurs, il faut le preciser par
# les arguments indx=T et/ou indy=T
# Si x ou y est un facteur, les noms de modalite sont les
# noms des modalites du facteur.
# La generation du tableau de burt se fait par aggregation de sous tableaux.
#_____
#
sym <- F
#_____
# Cas ou x a une seule colonne
#_____
if(is.null(nomidx) || (length(nomidx) == 1))
nomidx <- gennoml(x, mod = nomidx)
if((is.null(nomidx) || (length(nomidx) == 1)) & is.factor(x)) {
nomidx <- levels(x)
nbmodx <- length(nomidx)

```

```

}
if(missing(y)) {
  sym <- T
  y <- x
  nomidy <- nomidx
}
else
{
  if((is.null(nomidy) || (length(nomidy) == 1)) & is.factor(y))
  {
    nomidy <- levels(y)
    nj <- length(nomidy)
  }
  if(is.null(nomidy) || (length(nomidy) == 1))
  nomidy <- gennoml(y, mod = nomidy)
}
warnold <- options()$warn
options(warn = -1)
y <- design(y)
x <- design(x)
options(warn = warnold)
n1 <- length(nomidx)
n2 <- length(nomidy)
p1 <- ncol(x)
p2 <- ncol(y)
ni <- sapply(x, countlev)
nj <- sapply(y, countlev)
tab <- matrix(0, nrow = n1, ncol = n2)
nid <- 1
for(i in 1:p1)
{
  nif <- nid + ni[i] - 1
  njd <- 1
  for(j in 1:p2)
  {
    njf <- njd + nj[j] - 1
    tab[nid:nif, njd:njf] <- table(x[, i], y[, j])
    njd <- njd + nj[j]
  }
  nid <- nid + ni[i]
}
dimnames(tab) <- list(nomidx, nomidy)
attr(tab, "nbmod") <- list(ni, nj)
attr(tab, "sym") <- sym
tab

```

```

}
Fic.burt=burtR( Tab2 )
# Pour le tableau disjonctif complet on utilise la fonction
disjonctifR = fonction(y, nomid = NULL, linind = F)
{
# -----
# Generation d'indicatrices associees a une multipartition ou ensembles de
# variables qualitatives ; si les noms des modalites ne sont pas fournis,
# ils sont generes automatiquement
# Si l'option linind=T est choisie, la derniere indicatrice de chaque
# variable est supprimee pour obtenir des indicatrices lineairement indep.
# -----
nomy <- as.character(substitute(y))
##### - TESTS (nature de y) - #####
# Si y est unidimensionnel test=T
if (is.factor(y)) test<-T
if ( ( !is.factor(y) ) && ( is.null(dim(y)) ) ) {test<-T;y<-as.factor(y)}
if (is.null(dim(y))) {test<-T;y<-as.factor(y)}
if (is.data.frame(y)) test<-F
##### ----- #####
options(warn = -1)
nbmod <- if (test) countlev(y) else sapply(y, countlev)
nn <- cumsum(nbmod)
nr <- if(test) length(y) else nrow(y)
nom <- if(is.null(nomid) || (length(nomid) <= 1))
gennoml(y, mod = nomid, nomy = nomy) else nomid
x<-y
ind<-if (test) matrix(0,length(x),1) else matrix(0,nrow(x),1)
p<-if (test) 1 else length(x)
for (i in 1 :p)
{
lev <- if (test) levels(x) else levels(x[,i])
pp <- if (test) countlev(x) else countlev(x[,i])
d <- matrix(0,nr,pp)
for (j in 1 :pp)
d[,j] <- if (test) {as.numeric(x==lev[j])} else {as.numeric(x[,i]==lev[j])}
ind <- cbind(ind,d)
}
ind <- ind[,-1]
if(linind) {nom <- nom[ - nn];ind <- ind[ , - nn]}
nomi <- if( ( !is.null(dimnames(y)) && !is.null(dimnames(y)[[1]]) ) &&
length(dimnames(y)[[1]]) != 0) dimnames(y)[[1]] else (paste("i",
1 :nr, sep = ""))
if (dim(ind)[[2]] != length(nom))
stop("Problemes de codage ? (codes de 1 a k ?)\n")

```

```

dimnames(ind) <- list(nomi, nom)
attr(ind, "nbmod") <- nbmod
options(warn = 0)
if (test) contr<-levels(y) else {
liste<-"list("
for (i in 1 :length(nbmod))
{
if (i!=length(nbmod))
liste<-paste(liste,names(y)[i],"=levels(",nomy,"$",names(y)[i],")",sep="") else
liste<-paste(liste,names(y)[i],"=levels(",nomy,"$",names(y)[i],")",sep="")
}
contr <- eval(parse(text = liste), envir = sys.parent(2))
}
attr(ind,"contrasts")<-contr
attr(ind, "class") <- "matrix"
ind
}
Fic.disj = disjonctifR ( Tab2 )
# Eboulis des valeurs propres
eboulisR = fonction ( res, subpl = F, kmax = NULL, prct = T, titl = T, cum = F, new = F,
titre = NULL, ...)
{
#-----
# Représentation, de l'histogramme des indices de niveau en
# en classification hierarchique ou de l'histogramme des valeurs propres
# apres une acp ou une afc.
# si subpl=T, cet histogramme est trace en surimpression avec locator()
# Si kmax est precise, seul les kmax premiers batons sont representees
# si cum==T les pourcentages cumules remplacent les numeros de facteurs
#-----
cexold <- par("cex")
if(subpl)
new <- T
class <- !is.null(res$height)
fact <- !is.null(res$values)
ro2 <- !is.null(res$ro2)
if (!(class | fact | ro2)) stop(" \n Le premier argument doit etre le resultat\n
d'une acp, \nd'une afc, \nd'une sélection pas à pas en analyse discriminante ou\n
d'une classification hierarchique\n")
if(class) {
height <- rev(res$height)
if(is.null(titre))
titre <- "Histogramme des indices de niveau"
}
else {

```

```

if(fact) {
height <- res$values
if(is.null(titre))
titre <- "Eboulis des valeurs propres"
}
else {
height <- res$ro2
names(height) <- dimnames(res$x)[[2]][res$ind]
prct <- F
if(is.null(titre))
titre <-
"Rapports de corr\351lation des variables s\351lectionn\351es"
}
}
imax <- length(height)
if(!is.null(kmax))
imax <- min(kmax, imax)
if(prct)
height <- (100 * height)/sum(height)
height <- height[1 :imax]
nom <- if(is.null(names(height))) as.character(1 :length(height)) else
names(height)
nom2 <- NULL
if(cum) {
ndigit <- 1
if(imax > 10)
ndigit <- 0
nom2 <- as.character(round(cumsum(height), ndigit))
}
if(!subpl) {
xx <- barplot(height, ...)
mtext(nom, 1, 0.5, at = xx)
if(!is.null(nom2))
mtext(nom2, 1, 1.5, at = xx)
box()
if(titl)
title(titre)
}
else {
par(mfg = c(1, 1, 1, 1))
subplot(barplot(height, ...))
if(titl) {
cat("Emplacement du titre defini a la souris\n")
text(locator(1), titre, cex = 7/10 * cexold, adj = 0)
}
}

```

```
}  
invisible()  
}
```

---

# Bibliographie

---

- [1] *Cours du Dr Eugène Patrice Ndong Nguéma Data Mining, master de statistiques ENSP décembre 2006*
- [2] *Philippe Tassi, "Fondements statistiques " Collection Economica, vol2, pp 319-322.*
- [3] *cours du Pr Henry Gwet Données et Méthodologies, master de statistiques ENSP décembre 2006*
- [4] *Kengue J. (2002) Safou.Dacryodes edulis. Internatinal Centre for Underutilised Crops, Southampton, UK.*
- [5] *Nya Ngatchou J. et J.Kengue, (1986), Biologie florale et morphologie du safoutier ; présentation sommaire de l'arbre, Revue Science et Technique, série Sciences agricoles, 2(2), 47 - 65.*
- [6] *Kengue J. et J.Nya Ngatchou, (1991), Perspectives d'amélioration du safoutier, Actes du séminaire sous-régional sur la valorisation du safoutier 25 - 28 Nov. 1991, Brazzaville.*
- [7] *K.V.Mardia, J.T.Kent and J.M.Bibby, Multivariate Analysis, Academic Press, 213 - 276.*
- [8] *Alain Baccini et Philippe Besse, (2005), Data mining l'Exploration statistique, Laboratoire de Statistiques et de Probabilités - UMR CNRS C5583 Université Paul Sabatier - 31062 - Toulouse cedex 4.*
- [9] *S.H.C. du Toit, A.G.W.Steyn and R.H.Stumpf, Graphical Exploratory Data Analysis, Springer Texts in Statistics.*
- [10] *Philippe Bess ©1999 André Carlier and Mathieu Ros*

---

---

# Table des matières

---

---

---

# Table des figures

---

---

---

# Liste des tableaux

---