Chapitre 1

Introduction

L'analyse statistique multivariée consiste à analyser et comprendre des données de grande dimension. Nous supposons que nous avons un ensemble $\{x_i\}_{i=1,\cdots,n}$ de n observations d'un vecteur de variables X dans $\mathbb{R}^p$. Autrement dit, nous supposons que chaque observation x_i admet p dimensions :

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$$

et que c'est une valeur observée (ou réalisation) d'un vecteur de variables $X \in \mathbb{R}^p$. Le vecteur X est composé de p variables aléatoires :

$$X = (X_1, X_2, ..., X_p)$$

où X_j, pour $j = 1, \cdots, p$, est une variable aléatoire de dimension 1. Comment allons nous analyser ce type de données ? Avant de considérer la question de ce qu'on peut inférer à partir de ces données, on doit penser à comment regarder les données. Ceci implique des techniques descriptives. Les questions auxquelles nous pouvons répondre à l'aide d'analyses descriptives sont :

- Y a t'il certaines composantes de X qui sont plus dispersées que d'autres ?
- Y a t'il des éléments de X qui indiquent des sous-groupes dans les données ?
- Y a t'il des valeurs extrêmes et/ou aberrantes dans des données ?
- La distribution des données est-elle "normale" ?
- Y a t'il des combinaisons linéaires de faible dimension de X qui montrent des comportements "non-normaux" ?

Une difficulté des méthodes descriptives pour les données de grande dimension est le système de perception humain. Les nuages de points en deux dimensions sont faciles à comprendre et à interpréter. Avec les techniques de visualisation interactives modernes on a la possibilité de voir des rotations 3D en temps réel et ainsi percevoir aussi les données à 3 dimensions. Une technique de glissement [1] décrite par Härdle et Scott (1992) permet de matérialiser une 4ème dimension en représentant des contours 3D avec la 4ème dimension en niveau de couleur.

Un saut qualitatif dans les difficultés de représentation apparaît pour des dimensions supérieurs à 5, à moins que la structure de grande dimension ne puisse être projetée dans un espace de dimen-

1. sliding technic

sion plus faible. Certaines caratéristiques telles que des sous-groupes ou des valeurs aberrantes peuvent être détectées par des techniques d'analyses purement graphiques.

Dans le chapitre suivant, nous faisons quelques rappels importants d'algèbre linéaire. Dans le chapitre 3, nous introduisons l'analyse factorielle qui permet de projeter des données de grande dimension dans un espace de dimension plus faible. Nous en déduirons une technique classique : l'analyse en composantes principales. Dans le chapitre 4, nous étudierons un autre type d'analyse factorielle dont l'objectif est davantage un objectif de modélisation que de description : l'analyse en facteurs communs et spécifiques. Dans le chapitre 5, nous considérons un problème dans lequel on cherche des liens entre des variables (explicatives) continues et une variable (à expliquer) catégorielle et nous décrirons l'analyse factorielle discriminante. Puis dans le chapitre 6, nous nous intéresserons aux tableaux de données catégorielles et nous étudierons l'analyse des correspondances et l'analyse des correspondances multiples. En enfin dans le chapitre 7, nous nous tournons vers le problèmes de la classification supervisée qui permet de mettre en évidence des sous groupes dans les données. Pour conclure, dans le chapitre 8, nous mettrons en évidence que tous les problèmes évoqués peuvent être formalisés comme des problèmes d'inférence sur une ou plusieurs variables latentes.

Une partie des exemples de ce cours sont empruntés à Härdle et Simar (2007).

Chapitre 2

Rappels et compléments d'algèbre linéaire
Décompositions de matrices

2.1 Les projecteurs

La notion de projection est fondamentale en statistique. Par exemple la moyenne est une projection sur la droite des constantes. L'analyse en composante principale est basée sur des projecteurs de même que a régression linéaire.

2.1.1 Sous espaces supplémentaires et projecteurs

Soient F et G deux sous espaces vectoriels de E.

$$F + G = \{x + y | x \in F, y \in G\} \text{ et } F \times G = \{(, y) | x \in F, y \in G\}.$$

Définition 1 *On dit que F et G sont supplémentaires si $F \cap G = \emptyset$ et $F + G = E$.*
De façon équivalente, 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 espave vectoriel n'est pas unique.

Proposition 1 *Si F et G sont supplémentaires, les applications p et q de E dans E définies par*

$$\forall x \in E, x = p(x) + q(x) \text{ avec } p(x) \in F \text{ et } q(x) \in G$$

sont linéaires (on dit que ce sont des endomorphismes de E) et vérifient
[P1] $p^2 = p$; $q^2 = q$ (idempotence)
[P2] $poq = qop = 0$
[P3] $p + q = Id_E$
[P4] $Im(p) = F = Ker(q)$ et $Im(q) = F = Ker(p)$
On dit que p est la projection sur F parallèlement à G et que $q = Id_E - p$ est la projection sur G parallèlement à F.
On appelle projecteur dans un espace vectoriel E tout endomorphisme idempotent de E.

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

2.1.2 Exemple fondamental

Soient u et v de $\mathbb{R}^n$ muni du produit scalaire usuel, tels que

$$\langle u, v\rangle = v^T u = 1$$

Remarquons que puisque $\langle u, v\rangle = ||u||_2||v||_2 \cos(u, v)$, la condition précédente impose que l'angle vectoriel entre u et v est aigu. Considérons la matrice $n \times n$,

$$P = uv^T.$$

Cette matrice jouit des propriétés suivantes :

$$P^2 = uv^T uv^T = uv^T = P$$

et si $x \in Imu$, c'est à dire si $x = \alpha u$,

$$Px = uv^T(\alpha u) = \alpha uv^T u = \alpha u = x$$

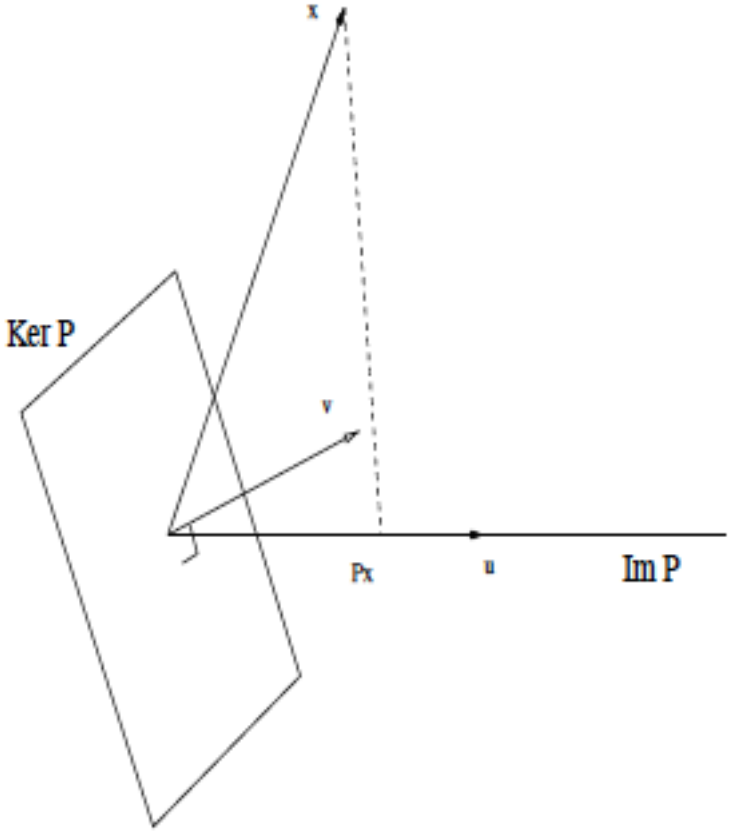
Mais si x est orthogonal à v, alors

$$Px = uv^T x = u(v^T x) = 0$$

L'image de P est donc Imu, le noyau de P est le sous espace vectoriel de dimension $n - 1$ orthogonal à v.

$$\mathbb{R}^n = Imu \oplus (Imv)^{\perp}.$$

P est donc la matrice de l'application linéaire "projection sur u parallèlement à $Imv^{\perp}$".



Si on choisit $v = u$ et $||u||_2 = 1$, le projecteur orthogonal s'écrit

$$P = uu^T.$$

De façon plus générale, soit F donné ainsi qu'une base $\{u_1, \cdots, u_r\}$ orthonormée de F. Soit $U = [u_1, \cdots, u_r]$ alors $U^T U = I_r$. La matrice

$$P = \sum_{i=1}^{r} u_i u_i^T = UU^T$$

est le projecteur orthogonal sur $F = ImU$. Le projecteur ($P^2 = P$) est orthogonal car $P = P^T$. En effet, la projection orthogonale est telle que pour tout vecteur quelconque Y de E, on cherche $\hat{Y} \in F$ tel que

$$(Y - \hat{Y}) \perp F$$

c'est à dire

$$\forall i, \quad u_i^T (Y - \hat{Y}) = 0$$

$$U^T (Y - \hat{Y}) = 0$$

D'oú

$$U^T Y = U^T \hat{Y}$$

Ceci signifie aussi que $\hat{Y}$ s'écrit comme une combinaison linéaire des éléments u_i,

$$\hat{Y} = c_1 u_1 + \cdots + c_k u_k = U \begin{bmatrix} c_1 \\ c_2 \\ \cdots \\ c_k \end{bmatrix}$$

On a donc

$$U^T Y = U^T U C$$

ce qui s'écrit aussi

$$(U^T U)^{-1} U^T Y = C$$

Et on remarque que $\hat{Y} = UC = U(U^T U)^{-1} U^T Y$ et on obtient que la matrice de projection est $U(U^T U)^{-1} U^T$.

Remarque - On reconnait les formules du modèle de régression linéaire pour des variables centrées (en prenant $U = X$, on a bien $C = var(X)^{-1} cov(X, Y)$).

Exercice Calculer la matrice de projection Q sur le sous espace de $\mathbb{R}^4$ engendré par les vecteurs $(1, 1, 0, 2)$ et $(-1, 0, 0, 1)$. Donner la projection de $x = (0, 2, 5, -1)$ sur le sous espace.

Exercice Calculer la projection de $v = (1, 1, 0)$ sur le plan $x + y - z = 0$.

2.2 Matrices carrées diagonalisables

Définition 2 *Une matrice carrée A d'ordre n est diagonalisable si elle est semblable à une matrice diagonale $\Lambda = diag(\lambda_1, \cdots, \lambda_n)$ ie qu'il existe une matrice inversible S telle que*

$$\Lambda = S^{-1}AS$$

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

Condition nécessaire et suffisante : Une condition nécessaire et suffisante pour que A carrée d'ordre n, soit diagonalisable est que ses n vecteurs propres soient linéairement indépendants.
Condition suffisante : Les vecteurs propres associés à des valeurs propres distinctes sont linéairement indépendants. Si toutes les valeurs propres de A sont distinctes, alors A est diagonalisable.

Décomposition spectrale de A diagonalisable
Soit A diagonalisable telle que $A = S\Lambda S^{-1}$. Notons u_j la $j$ème colonne de S et v_j^T la $j$ème ligne de S^{-1}, associés à λ_j. La décomposition spectrale de A s'écrit

$$A = \sum_{j=1}^{n} \lambda_j u_j v_j^T$$

Le vecteur v_j est le vecteur propre de A^T associé à $\bar{\lambda}_j$ et $v_j^T u_i = 0$ si $j \neq i$. Ceci signifie que les vecteurs propres distincts de A et A^T sont orthogonaux.

- Une matrice symétrique et réelle est diagonalisable et on a $A = S\Lambda S^T$.
- Une matrice symétrique et réelle est (semi) définie positive si et seulement si toutes ses valeurs propres sont positives (non négatives).

2.3 Décomposition en valeurs singulières

Pour une matrice rectangulaire, la notion de valeur propre n'a pas de sens. Néanmoins, les matrices carrées A^TA et AA^T sont symétriques semi définies positives. De plus,

$$rang(A) = rang(AA^T) = rang(A^TA) = r$$

et les r valeurs propres non mulles (positives) de A^TA et AA^T sont identiques.

Définition 3 *On appelle valeurs singulières de A les racines carrées des valeurs propres non nulles et A^TA ou de AA^T.*

$$\mu_i = \sqrt{\lambda_i(A^TA)} = \sqrt{\lambda_i(AA^T)}$$

Soit A $m \times n$ telle que $rang(A) = r$. Alors

$$A = U\Lambda_r^{1/2}V^T = \sum_{j=1}^{r} \mu_j u_j v_j^T$$

avec

- $U = [u_1, \cdots, u_r]$ unitaire $m \times r$ est telle que u_j est le vecteur propre de AA^T associé à la valeur propre non nulle λ_i.
- $V = [v_1, \cdots, v_r]$ unitaire $n \times r$ est telle que v_j est le vecteur propre de A^TA associé à la valeur propre non nulle λ_i.
- $\Lambda_r = diag(\lambda_1, \cdots, \lambda_r)$ et $\Lambda_r^{1/2} = diag(\mu_1, \cdots, \mu_r)$ où $\mu_j = \lambda_j^{1/2}$ est la $j$ème valeur singulière de A.

Remarques/résultats importantes -
- Dans la pratique, le nombre de valeurs singulières non nulles fourni le rang de la matrice.
- Dans le calcul de U et de V , on ne calcule les vecteurs propres de AA^T ou de A^TA que pour celle de ces matrices de plus petite dimension, les vecteurs propres de l'autre se déduisent par des "formules de transition" (2.2) et (**??**).

$$U = AV\Lambda_r^{-1/2} \tag{2.1}$$

avec $\Lambda_r^{-1/2} = (\Lambda_r^{1/2})^{-1} = diag(1/\mu_1, \cdots, 1/\mu_r)$

$$V = A^TU\Lambda_r^{-1/2} \tag{2.2}$$

- La décomposition en valeurs singulières donne

$$A^TA = V\Lambda_rV^T = \sum_{i=1}^{r} \mu_i^2 v_i v_i^T$$

$$AA^T = U\Lambda_rU^T = \sum_{i=1}^{r} \mu_i^2 u_i u_i^T$$

- Il y a d'importantes projections orthogonales associées à la décomposition en valeurs singulières. Soit A supposée de rang r et $A = U\Lambda_r^{1/2}V^T = P\Lambda Q^T$, la SVD de A. Rappelons que les partitions des colonnes de P et Q

$$P = [U|\tilde{U}], \quad Q = [V|\tilde{V}]$$

avec U les r premières colonnes de P et $\tilde{U}$ les suivantes (idem pour V et Q)
  - VV^T = projection orthogonale sur $\{KerA\}^\perp = ImA^T$
  - $\tilde{V}\tilde{V}^T$ = projection orthogonale sur $KerA$
  - UU^T = projection orthogonale sur ImA
  - $\tilde{U}\tilde{U}^T$ = projection orthogonale sur $\{ImA\}^\perp = KerA^T$
- On peut montrer que l'approximation d'une matrice A de rang p par une matrice de B de rang $q < p$ est donnée par la décomposition en valeurs singulière $B = U\tilde{\Lambda}V^T$ avec $\tilde{\Lambda}$ une matrice diagonale qui contient les q plus grandes valeurs singulières de A.

2.4 Les projecteurs M-orthogonaux

En statistique on est souvent amené à définir des produits scalaires différents du produit scalaire usuel et basés sur des métriques M, où M est une matrice symétrique définie positive, différentes de l'identité.

$$\langle x, y\rangle_M = y^TMx \text{ et } ||x||_M = x^TMx.$$

Définition 4 *Soit l'espace vectoriel Euclidien $\mathbb{E} = \mathbb{R}^m$ mini d'un M produit scalaire et soit $\mathbb{E}_1$ un sous espace vectoriel de $\mathbb{E}$ tel que $\mathbb{E} = \mathbb{E}_1 \oplus \mathbb{E}_1^\perp$ où $\mathbb{E}_1^\perp = \{y \in \mathbb{E} | \langle y, x \rangle_M = 0, x \in \mathbb{E}_1\}$. Pour tout x de $\mathbb{E}$ la décomposition*

$$x = x_1 + y_1,\ x \in \mathbb{E}_1,\ y \in \mathbb{E}_1^\perp$$

est unique. P est un projecteur M-orthogonal sur $\mathbb{E}_1$ si et seulement si

$$Px = x_1(I - P)x = y_1$$

La notion de M-orthogonalité est liée à une notion de symétrie particulière, la M-symétrie. La symétrie usuelle correspond au cas où M est l'identité.

Définition 5 *Une matrice $A \in \mathbb{R}^{m \times m}$ est M-symétrique si*

$$MA = A^T M$$

c'est à dire que MA est symétrique.

Proposition 2 *Une projecteur P est un projecteur M-orthogonal si et seulement si P est M-symétrique.*

Preuve : Soit P un projecteur ($P^2 = P$) sur $\mathbb{E}_1$ tel que

$$\forall x, y \in \mathbb{E},\ Px \in \mathbb{E}_1,\ (I - P)y \in \mathbb{E}_1^\perp \text{ au sens de } M.$$

c'est à dire que $x^T P^T M(I - P)y = 0 \equiv P^T M(I - P) = 0 \equiv P^T M = (P^2)^T M = P^T MP$.
Puisque M est symétrique, $P^T M$ est aussi symétrique, $P^T M = MP$. $\diamond$

Chapitre 3

Analyse en Composantes Principales

3.1 Introduction

L'objectif de ce chapitre est d'étudier les méthodes classiquement utilisées pour décrire et visualiser des données multivariées issues de variables continues : l'analyse en composantes principales et le positionnement multidimensionnel. Ces techniques d'analyse descriptive seront utilisées, notamment, pour visualiser les données dans un sous espace représentatif, pour détecter des groupes d'individus et/ou de variables, des valeurs extrèmes ou abberrantes ou pour aider au choix de variables. Ces méthodes permettent aussi de répondre à des questions du type : quels individus se ressemblent du point de vue de l'ensemble des variables ? ou réciproquement quelles variables sont semblables du point de vue des l'ensemble des individus ?

L'analyse en composantes principales est un outil de réduction de dimension qui permet de retirer la redondance ou la duplicité dans un ensemble de variables corrélées. L'ensemble initial est alors représenté par un ensemble réduit de variables dérivées des variables observées. Ces facteurs sont, en théorie, indépendants les uns des autres et on peut les classer par ordre d'importance.

Soit $\{X_1, \cdots, X_p\}$ un ensemble de p variables observées sur n individus indépendants. On notera

$$\mathbf{x} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \cdots & & \cdots \\ x_{np} & \cdots & x_{np} \end{pmatrix}$$

le tableau des n observations des p variables. Typiquement, n est grand devant p. Pour tout $j \in \{1, \cdots, p\}$, $x_j \in \mathbb{R}^n$. Chaque ligne du tableau représente un individu et chaque colonne une variable. Chaque individu est un point de l'espace $\mathbb{R}^p$. On dira que $\mathbb{R}^p$ est l'espace des variables et $\mathbb{R}^p$ l'espace des individus.

Par exemple, dans le tableau ci-dessous la première colonne (Modèle) est l'identifiant et on observe deux variables, la puissance de la voiture et son prix.

Modèle	Puissance	Prix
Alfasud TI	79	30570
Audi 100	85	39990
Simca 1300	68	29600
Citroen GS Club	59	28250
Fiat 132	98	34900
Lancia Beta	82	35480
Peugeot 504	79	32300
Renault 16 TL	55	32000
Renault 30	128	47700
Toyota Corolla	55	26540
Alfetta-1.66	109	42395
Princess-1800	82	33990
Datsun-200L	115	43980
Taunus-2000	98	35010
Rancho	80	39450
Mazda-9295	83	27900
Opel-Rekord	100	32700
Lada-1300	68	22100

Quand on n'observe que deux variables, la représentation des individus est directe : on représente les individus dans le plan de $\mathbb{R}^2$, chaque axe représentant une variable. L'objectif de l'analyse en composantes principales est de représenter les individus quand $p > 2$. L'idée est la suivante. Supposons dans un premier temps que les individus soient en fait concentrés dans un plan de $\mathbb{R}^p$. La solution la plus simple consiste à faire un changement de base où les deux premiers axes sont dans le plan et les autres leurs sont orthogonaux (et les coordonnées des individus sur les axes 3 à p axes seront nulles). Considérons maintenant un nuage de points qui est presque concentré sur un plan. En pratique, on cherche un sous espace vectoriel dans lequel la dispersion entre les observations est la mieux représentée. On cherche aussi à préserver au mieux les distances entre les individus. On peut faire l'analogie avec une photographie. Si on photographie un objet en 3 dimensions (par exemple un poisson), on va chercher un plan tel qu'on reconnaisse aisement que c'est un poisson, c'est à dire un plan dans lequel les informations importantes sont restituées au mieux. Dans la figure 3.1 l'image représentant le poisson de profil (plus grande dispersion) restitue davantage d'information que celle du poisson de face (moins de dispersion).

D'un point de vue plus "mathématique", l'ACP correspond à l'approximation d'une matrice $n \times p$ par une matrice de même dimension mais de rang $q < p$, q étant souvent de petite valeur 2, 3 pour la construction de graphiques facilement compréhensibles. Plus précisement, les objectifs poursuivis par l'ACP sont

- la représentation graphique "optimale" des individus en minimisant les déformations du nuage des points, dans un sous espace de dimansion $q < p$ (autrement dit on cherche à préserver les distances entre individus) ;
- la représantation graphique des variables dans un sous espace E_q en explicitant "au mieux" les liaisons entre ces variables ;
- la réduction de la dimansion (compréssion), ou approxiation du tableau de données X par une matrice de rang $q < p$.

FIGURE 3.1 – Le poisson clown.

On peut construire l'ACP de plusieurs façons. L'approche la plus classique (en France) est l'approche géométrique.

3.2 ACP par projection : approche géométrique

En ACP, on travaille toujours sur les données centrées. Notons $\tilde{x}_i$ et $\tilde{\mathbf{x}}$ les individus centrés :

$$\tilde{x}_i = x_i - \bar{x}_i, \tilde{\mathbf{x}} = \begin{pmatrix} x_1 - \bar{x}_1 \\ \dots \\ x_n - \bar{x}_n \end{pmatrix}$$

La moyenne empirique $\bar{x}$ est parfois appelée centre de gravité.

La dispersion d ?un nuage de points unidimensionnel par rapport à sa moyenne se mesure par la variance. Dans le cas multidimensionnel, la dispersion du nuage par rapport à son barycentrese mesure par l'inertie.

Définition 6 *L'inertie des individus est donnée par la quantité*

$$I = \frac{1}{n} \sum_{i=1}^{n} ||\tilde{x}_i||^2$$

On remarque que l'*inertie* est définie comme la somme des distances au carré des points à leur centre de gravité. Dans le cas où les variables dont quantitatives, c'est aussi la somme des variances empiriques de chacune des variables, c'est à dire la trace de la matrice de variance-covariance empirique $\hat{\Sigma}$. En effet,

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i^T \tilde{x}_i = \frac{1}{n} \tilde{\mathbf{x}}^T \tilde{\mathbf{x}}, \hat{\Sigma}_{jk} = \frac{1}{n} \sum_{i=1}^{n} \tilde{x}_{ij} \tilde{x}_{ik}$$

L'inertie est une quantité réelle qui mesure la dispersion des individus dans l'espace à p dimensions.

Soit P un projecteur de $\mathbb{R}^p$. Par abus, on notera également P la matrice associée à P dans la base canonique. La projection d'un vecteur x_i sera

$$P(x_i) = x_i P^T, \quad XP^T = \begin{pmatrix} x_1 P^T \\ \dots \\ x_n P^T \end{pmatrix}$$

Soit E un sous-espace de $\mathbb{R}^p$ et P_E le projecteur orthogonal sur E, on note I_E l'inertie des individus projetés :

$$I_E = \frac{1}{n}\sum_{i=1}^{n} ||P_E(\tilde{x}_i)||^2 = \frac{1}{n}\sum_{i=1}^{n} ||\tilde{x}_i||^2 - \frac{1}{n}\sum_{i=1}^{n} ||\tilde{x}_i - P_E(\tilde{x}_i)||^2$$

par Pythagore. L'inertie I_E est donc également une mesure de la dispersion des individus après projection sur E. Il est facile de vérifier que

$$I_E = Tr(P_E \hat{\Sigma} P_E)$$

Soit $u_1, \cdots, u_q$ une base orthogonale de E. Alors $P_E = UU^T$ où U est la matrice rectangulaire formée des vecteurs U_i en colonne : $U = [u_1, \cdots, u_q]$. Donc, la trace étant invariante pas changement de base,

$$I_E = Tr(P_E \hat{\Sigma} P_E) = Tr(U^T \hat{\Sigma} U) = \sum_{i=1}^{q} u_i^T \hat{\Sigma} u_i$$

.

Raisonnons dans un premier temps avec un seul axe de projection u_1, ie $q = 1$. La projection d'un individu observé $\tilde{x}_i \in \mathbb{R}^p$ sur l'axe u est définie par

$$P_u(x_i) = x_i^T \frac{u}{||u_1||}$$

Et on cherche l'axe u^* qui conduit à la projection qui conserve au mieux les distances entre individus :

$$u^* = \min_{u \in \mathbb{R}^p, ||u||=1} \sum_{i=1}^{n} ||\tilde{x}_i - P_u(\tilde{x}_i)||^2 \tag{3.1}$$

avec $\tilde{\mathbf{x}}$ le nuage de points centré (et éventuellement réduit) et $\tilde{x}_i$ le $i$ème individu correspondant. Par le théorème de Pythagore, on sait que $||\tilde{x}_i - P_u(\tilde{x}_i)||^2 = ||\tilde{x}_i||^2 - ||P_u(\tilde{x}_i)||^2$, ainsi le problème de l'équation (3.1) est équivalent à

$$u^* = \max_{u \in \mathbb{R}^p, ||u||=1} \sum_{i=1}^{n} ||P_u(\tilde{x}_i)||^2 \tag{3.2}$$

soit encore en utilisant la définition de l'opérateur de projection :

$$u^* = \max_{u \in \mathbb{R}^p, ||u||=1} \sum_{i=1}^{n} u^T \tilde{\mathbf{x}}^T \tilde{\mathbf{x}} u$$

On remarque que la variance empirique de $P_u(\tilde{\mathbf{x}})$ vaut

$$\frac{1}{n} P_u(\tilde{\mathbf{x}})^T P_u(\tilde{\mathbf{x}}) = u^T \cdot \underbrace{\frac{1}{n} \tilde{\mathbf{x}}^T \tilde{\mathbf{x}}}_{\hat{\Sigma}} \cdot u$$

où $\hat{\Sigma}$ est la matrice de covariance empirique de $\tilde{\mathbf{x}}$. Ainsi, pour le premier vecteur propre, on cherche un vecteur unitaire u^* tel que

$$u^* = arg \max_{\{u \in \mathbb{R}^n, u^T u = 1\}} u^T \hat{\Sigma} u \tag{3.3}$$

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, (3.3) est équivalent à

$$(u^*, \lambda_1) = arg \max_{\{u \in \mathbb{R}^n, \lambda \in \mathbb{R}\}} u^T \hat{\Sigma} u - \lambda (u^T u - 1)$$

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

$$\frac{\partial \left(u^T \hat{\Sigma} u - \lambda (u^T u - 1) \right)}{\partial u} = 2\Sigma u - \lambda u$$

$$\frac{\partial \left(u^T \hat{\Sigma} u - \lambda (u^T u - 1) \right)}{\partial \lambda} = u^T u - 1$$

Si on remarque maintenant que

$$\max_{\{u \in \mathbb{R}^n, u^T u = 1\}} u^T \hat{\Sigma} u = \max_{\{u \in \mathbb{R}^n, u^T u = 1\}} u^T \lambda u = \max_{\{u \in \mathbb{R}^n, u^T u = 1\}} \lambda$$

on a que le premier axe factoriel u^* est associé à la plus grande valeur propre de $\hat{\Sigma} u$.

Plus généralement, la maximisation de l'inertie I_E sur toutes les familles de q vecteurs orthogonaux est réalisée en choisissant les q vecteurs associés aux q plus grandes valeurs propres de $\hat{\Sigma}$ et on a les théorèmes suivants.

Théorème 1 *L'espace de dimension q d'inertie maximale est engendré par les q vecteurs propres associés aux q plus grandes valeurs propres (si des valeurs propres sont égales il n'y a pas unicité).*

Théorème 2 *Les composantes principales sont données par la transformation linéaire* $Y = U^T(X - E(X))$ *où* $\hat{\Sigma} = Var(X) = U \Lambda U^T$. *De plus on a :*

$$E(Y_j) = 0, \ \forall j = 1, \cdots, p$$

$$Var(Y_j) = \lambda_j, \ \forall j = 1, \cdots, p$$

$$Cov(Y_j, Y_k) = 0, \ \forall j, k = 1, \cdots, p$$

3.3 Représentations graphiques et aide à l'interprétation

L'analyse en composantes principales est principalement utilisée pour donner une représentation graphique des individus et des variables.

3.3.1 Les individus

En pratique, on projette orthogonalement les observations $\tilde{\mathbf{x}}$ sur les plans factoriels. Les coordonnées de $\mathbf{x}_i - \bar{\mathbf{x}}$ sur le sous espace de dimension q sont les q premiers éléments de la matrice $C = U\Lambda^{1/2}$. Voir l'exemple ci-dessous. Les graphiques obtenus permettent de représenter au mieux les distances euclidiennes inter-individus.

La qualité globale des représentations est mesurée par la *part de dispersion expliquée* ou la *la portion d'inertie expliquée* :

$$r_Q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$$

Tandis que la qualité de la représentation de chaque point est donnée par

$$cs_i^2 = \frac{\sum_{k=1}^{q} d(O, y_i)_k^2}{\sum_{j=1}^{p} d(O, y_i)_k^2}$$

où $d(O, y_i)_k = c_{iK}$

La contribution de chaque individu à l'inertie du nuage permet de détecter les observations les plus influentes et éventuellement aberrantes.

$$\gamma_i = \frac{\sum_{j=1}^{p} c_{ij}^2}{\sum_{j=1}^{p} \lambda_j}$$

Si la contribution d'un individu à un ou plusieurs axes est beaucoup plus importante que celle des autres il faut vérifier si cet individu n'est pas abbérant.

On peut projeter des individus supplémentaires $\mathbf{s}$ sur un sous espace factoriel en calculant ses coordonnées :

$$U^T(\mathbf{s} - \bar{\mathbf{x}})$$

Ici U joue le rôle d'une matrice de changement de base.

3.3.2 Les variables

La projection des variables sur les plans factoriels peuvent aider à l'interprétation des composantes. Cette représentation des variables peut s'interpréter comme le positionnement, pour chaque variable, d'un individu type, pour lequel les autres variables auraient leur valeur moyenne et la variable considérée serait amplifiée. Les graphiques obtenus permettent de représenter "au mieux" les corrélations entre les variables et, si celles ci ne sont pas réduites, leurs variances. On obtient le cercle des corrélations par projection orthogonale sur le sous espace factoriel E_q. La coordonnée de la variable x_j sur u_k est donnée par

$$\sqrt{\lambda} u_{jk}$$

La qualité de la représentation de chaque x_j est mesurée par

$$\frac{\sum_{j=1}^{q} \lambda_j v_{jk}^2}{\sum_{j=1}^{p} \lambda_j v_{jk}^2}$$

3.4 Exemple

A titre d'exemple, on considère un jeu de données établissant la composition du lait de 25 espèces de mammifères. On mesure 5 variables : la teneur en protéines, en lactose, en graisse, en eau et en minéraux. On obtient pour les matrices U et Λ suivantes.

$$U = \begin{pmatrix} 0.76 & -0.16 & -0.57 & -0.25 & -0.01 \\ -0.16 & 0.85 & -0.27 & -0.39 & -0.14 \\ -0.62 & -0.44 & -0.55 & -0.34 & 0.01 \\ 0.09 & -0.18 & 0.54 & -0.82 & 0.04 \\ -0.01 & 0.13 & -0.06 & -0.02 & 0.99 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 282.1 & 0 & 0 & 0 & 0 \\ 0 & 8.1 & 0 & 0 & 0 \\ 0 & 0 & 1.2 & 0 & 0 \\ 0 & 0 & 0 & 0.3 & 0 \\ 0 & 0 & 0 & 0 & 0.1 \end{pmatrix}$$

En première approximation, on peut dire que les composantes principales correspondent dans l'ordre à

- la proportion d'eau sur la proportion de graisse
- la teneur en protéines
- la proportion de lactose sur celle d'eau et de graisse
- la teneur en lactose
- la teneur en sel minéraux

Le fait que la première valeur propre soit grande devant les autres signifie que les individus se démarquent surtout par la proportion d'eau par rapport à la graisse dans leur lait. La figure 3.2 montre, dans le premier plan factoriel, les graphes des variables et des individus pour l'ACP non réduite. Ce plan explique 99.5% de la variance. On observe sur le graphe de projection des variables que l'eau est le composant le plus important suivi par la matière grasse dans le composant du lait. Associé au graphe des inidividus, on peut voir, par exemple, que le dauphin et le phoque ont des laits plus gras que les autres mammifères. Le graphe permet de visaulisé que les variables qui contribuent fortement au premier axe factoriel sont la matière grasse et l'eau. Le deuxième axe factoriel apporte peu d'information supplémentaire ; ce sont essentiellement les protéines qui contribuent à cet axe.

Le plus souvent, il est préférable d'interpréter une ACP réduite dans laquelle chaque variable va avoir la même contribution. Le résultat est alors indépendant des unités utilisées. Dans le cas de l'exemple les différents composants du lait sont mesurées dans les même unités. On peut alors préférer ne pas normaliser car les grandeurs relatives des variables sont importantes.

Si on normalise les données, on obtient pour des matrices U et D analogues à celes du cas non normalisé,

$$U = \begin{pmatrix} 0.47 & 0.35 & 0.37 & 0.11 & 0.71 \\ -0.47 & 0.32 & 0.15 & -0.79 & 0.19 \\ -0.45 & -0.48 & -0.31 & 0.18 & 0.67 \\ 0.48 & 0.06 & -0.78 & -0.38 & 0.11 \\ -0.35 & 0.74 & -0.38 & 0.43 & -0.00 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 3.88 & 0 & 0 & 0 & 0 \\ 0 & 0.89 & 0 & 0 & 0 \\ 0 & 0 & 0.13 & 0 & 0 \\ 0 & 0 & 0 & 0.10 & 0 \\ 0 & 0 & 0 & 0 & 0.01 \end{pmatrix}$$

La figure 3.3 montre, dans le premier plan factoriel, les graphes des variables et des individus pour l'ACP réduite. Le graphe des variables est aussi appelé cercle des corrélations. Le premier plan factoriel restitue 95.3% de la variance. Le cercle des corrélations permet de dire que les laits à forte teneur en matière grasse ou protéines sont généralement à faible teneur en lactose et eau car ces variables sont opposées sur le graphe. Ce sont ces variables qui contribuent au premier axe factoriel. Le second axe oppose les laits riches en protéines et minéraux aux laits riches en matières grasses. On remarque à l'aide du graphe des individus que les animaux qui ont un lait riche en eau sont surtout des animaux de régions chaudes.

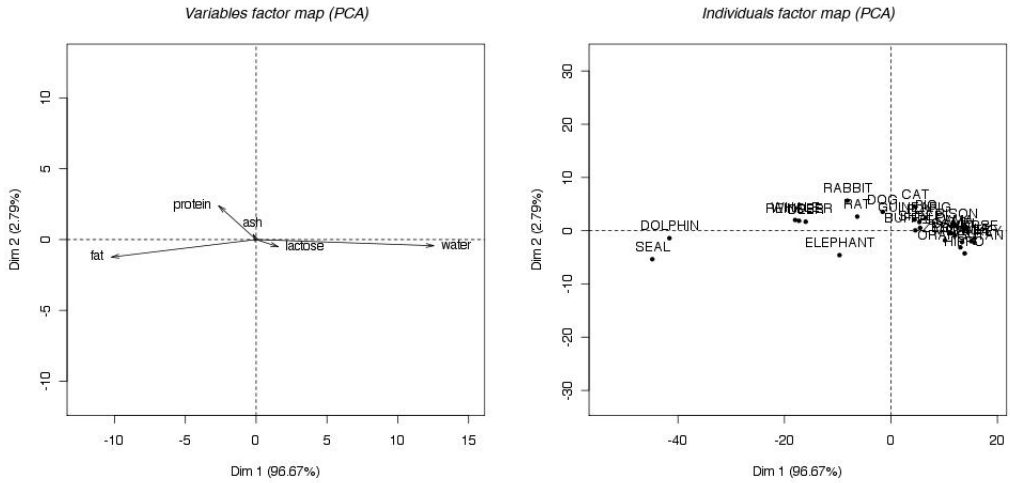


FIGURE 3.2 – Composition du lait - Cercle des corrélations (à gauche) et graphe des individus (à droite) sur le premier plan principal de l'ACP non réduite

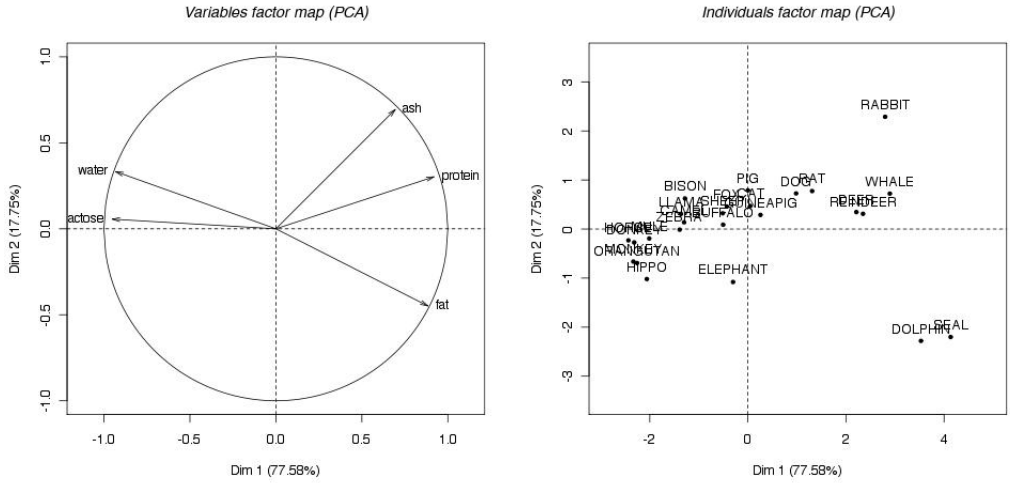


FIGURE 3.3 – Composition du lait - Cercle des corrélations (à gauche) et graphe des individus (à droite) sur le premier plan principal de l'ACP non réduite

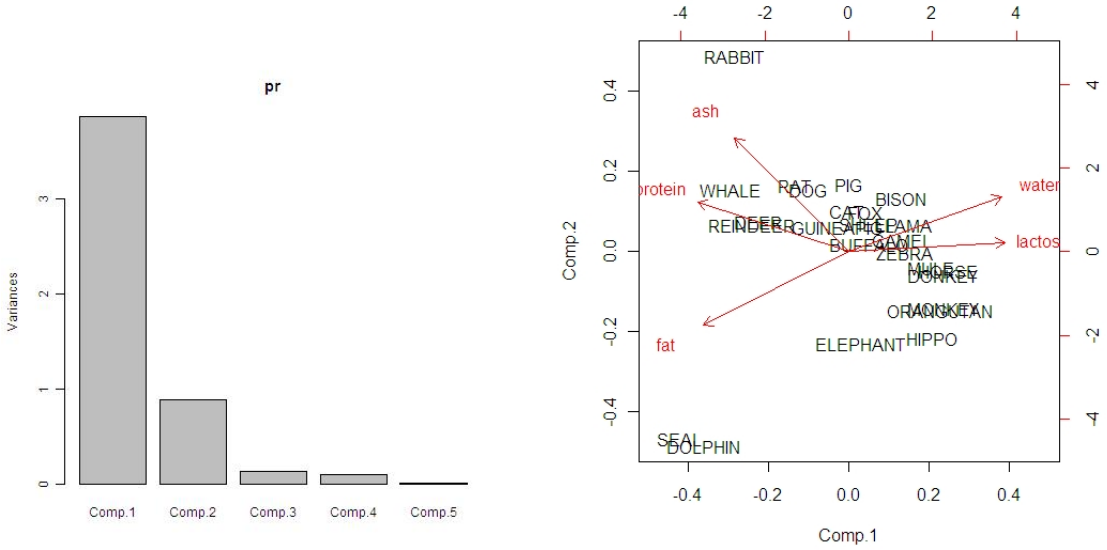


FIGURE 3.4 – Composition du lait - Ebouli des valeurs propres (à gauche) et représentation simultannée sur le premier plan principal de l'ACP (à droite)

3.5 Propriétés asymptotiques des estimateurs de composantes principales

En pratique l'ACP est réalisée à partir de données. On manipule donc des estimateurs. Il est utile de connaitre leurs propriétés.

Théorème 3 *Soit* $\Sigma > 0$ *ayant des valeurs propres distinctes et soit* $\hat{\Sigma} \sim n^{-1}W_p(\Sigma, n)$ *tels que* $\Sigma = \Gamma\Lambda\Gamma^T$ *et* $\hat{\Sigma} = \hat{\Gamma}\hat{\Lambda}\hat{\Gamma}$. *Alors*
(a) $\sqrt{n}(\hat{\lambda} - \lambda) \to_d \mathcal{N}_p(0, 2\Lambda^2)$,
avec $\hat{\lambda} = (\hat{\lambda}_1, \cdots, \hat{\lambda}_p)^T$ *et* $\lambda = (\lambda_1, \cdots, \lambda_p)^T$ *sont les diagonales de* $\hat{\Lambda}$ *et* Λ.
(b) $\sqrt{n}(g_j - \gamma_j) \to_d \mathcal{N}_p(0, \mu_j)$,
avec $\mu_j = \lambda_j \sum_{k \neq j} \frac{\lambda_k}{(\lambda_k - \lambda_j)^2} \gamma_k \gamma_k^T$.
(c) les éléments de $\hat{\lambda}$ *sont asymptotiquement indépendants de ceux de* Γ.

$W_p(\Sigma, n)$ est la loi de Wishart de variance Σ à n degrés de liberté. C'est une généralisation de la loi du khi pour les matrices aléatoires.

Comme $n\hat{\Sigma} \sim W - p(\Sigma, n-1)$ si $X_1, = \cdots, X_n$ sont distribuées suivant une loi de Gauss de moyenne μ et de variance Σ, on déduit du théorème que

$$\sqrt{n-1}(\hat{\lambda}_j - \lambda_j) \to \mathcal{N}(0, 2\lambda_j^2),\ j = 1, \cdots, p$$

En appliquant une transformation log, on obtient par la delta méthode,

$$\sqrt{n-1}(\log(\hat{\lambda}_j) - \log(\lambda_j)) \to \mathcal{N}(0, 2),\ j = 1, \cdots, p$$

et on peut alors écire un intervalle de confiance pour $\log(\lambda_j)$.

3.6 ACP par minimisation de l'erreur

On peut voir l'analyse en composantes principales comme un outil de synthèse d'information. L'idée est alors de chercher des facteurs latents sur lesquels se concentre l'information. Les facteurs latents jouent le même rôle que les composantes principales U. On peut alors écrire pour les variables centrées

$$\tilde{\mathbf{X}} = \sum_{j=k}^{q} c_q U_q + \epsilon$$

Dans le cas de l'ACP, on suppose que ϵ est un vecteur aléatoire gaussien centré dont les composantes sont indépendantes et de même variance : $\epsilon \sim \mathcal{N}(0, \sigma I)$. On a à faire à un modèle linéaire on peut donc réaliser l'inférence des paramètres inconnus z et U par minimisation de la variance des résidus. C'est à dire qu'on cherche les matrices $\mathbf{c}^*$ et $\mathbf{U}^*$ telles que

$$\begin{aligned}(\mathbf{c}^*, \mathbf{U}^*) &= arg \min_{\{(c,U)\in\mathbb{R}^q\times\mathbb{R}^q, \mathbf{U}\mathbf{U}^T = Id\}} Var\left(\mathbf{X} - \sum_{k=1}^{q} c_k U_k\right) \\ &= arg \min_{\{(c,U)\in\mathbb{R}^q\times\mathbb{R}^q, \mathbf{U}\mathbf{U}^T = Id\}} ||\mathbf{x} - \sum_{j=1}^{q} c_j U_j||^2 \end{aligned} \tag{3.4}$$

En pratique, on ne sait pas calculer cette variance. On l'estime à partir des observations. Et on montre que la solution unique est donnée par les composantes principales $\hat{U}$ et les axes principaux $\hat{z}$, vecteurs propres de la matrice de variance-covariance.

3.7 Changement de métrique dans l'espace des individus et poids sur les individus

Supposons maintenant que la mesure adéquate entre les individus n'est plus la distance euclidienne mais doive être basée sur une norme $||x||_M^2 = x^T M x$ où M est une matrice symétrique définie positive. La métrique M renormalise correctement les individus et il faut la prendre en compte dans le calcul d'inertie. Ceci est automatique si on considère la matrice des individus $\mathbf{x}' = \mathbf{x}M^{1/2}$. Dans la représentation des individus sur les axes factoriels c'est la nouvelle distance qui est approchée.

De manière analogue, si on veut donner des poids différents aux individus dans le calcul de l'inertie, on peut introduire une matrice de poids D qui est une matrice diagonale contenant les poids : $\mathbf{x}' = D^{1/2}\mathbf{x}M^{1/2}$.

Dans ce cas, on formule le problème (3.4) ainsi :

$$(\mathbf{c}^*, \mathbf{U}^*) = arg \min_{\{(c,U)\in\mathbb{R}^q\times\mathbb{R}^q, \mathbf{U}\mathbf{U}^T=Id\}} ||\mathbf{x} - \sum_{k=1}^{q} c_k U_k||^2_{(M,D)}$$

Si l'espace est euclidien, par définition

$$||\mathbf{x} - \sum_{j=1}^{q} c_j U_j||^2_{(M,D)} = D(\mathbf{x} - \sum_{j=1}^{q} c_j U_j)^T M(\mathbf{x} - \sum_{k=1}^{q} c_k U_k)$$

La solution est donnée par

$$\sum_{k=1}^{q} c_k U_k = \sum_{k=1}^{q} \lambda_j^{1/2} u_k v_k^T$$

avec U et V des matrices unitaires. C'est la décomposition en valeurs singulières de la matrice des données centrées réduites. Les vecteurs u_k sont les vecteurs propres de la matrice de covariance $\mathbf{x}M\mathbf{x}^T D$, les valeurs propres étant rangées par ordre décroissant. Tandis que les vecteurs v_k sont les vecteurs propres de $\mathbf{x}^T M\mathbf{x}D$ correspondant aux mêmes valeurs propres. Ils sont correspondant aux *axes principaux.*

A partir de V_q, matrice construite à partir des q premiers vecteurs v_k, on construit la matrice de projection $P_q = V_q V_q^T M$.

Le choix de la métrique M et/ou de la matrice de pondération D a un impact sur les résultats et notamment sur les projections des individus sur les plans factoriels. Certaines métriques permettent par exemple de mettre en évidence les individus atypiques (voir TD). Le plus souvent, on choisi $D = \frac{1}{n}\mathbf{I}$ et $M = \mathbf{I}$ avec $\mathbf{I}$ la matrice identité. C'est à dire qu'on donné le même point à chaque individu et qu'on ne priviligie aucune variable.