

Analyse en Composantes Principales

STID - 2A

Maxime FRANCOISE

2020 – 2021

- **Analyse en Composantes Principales (ACP)**
 - Composantes principales
 - Repère factoriel
 - Visualisation des données
 - Compression, débruitage
 - Utilisation de l'outil logiciel

Cadre d'utilisation de l'ACP

- **Tableau de données brutes** $n \times p$ constitué uniquement de variables **quantitatives** et tel que $n > p$.
- **Individu** i assimilé à un **point de \mathbb{R}^p** de coordonnées $(x_i^{(1)}, \dots, x_i^{(p)})$ (lecture du tableau en lignes).
- **Variable** $X^{(j)}$ assimilée à **point de \mathbb{R}^n** de coordonnées $(x_1^{(j)}, \dots, x_n^{(j)})$ (lecture du tableau en colonnes).

Individu \ Variable	Variable				
	$X^{(1)}$...	$X^{(j)}$...	$X^{(p)}$
1	$x_1^{(1)}$...	$x_1^{(j)}$...	$x_1^{(p)}$
2	$x_2^{(1)}$...	$x_2^{(j)}$...	$x_2^{(p)}$
...
n	$x_n^{(1)}$...	$x_n^{(j)}$...	$x_n^{(p)}$

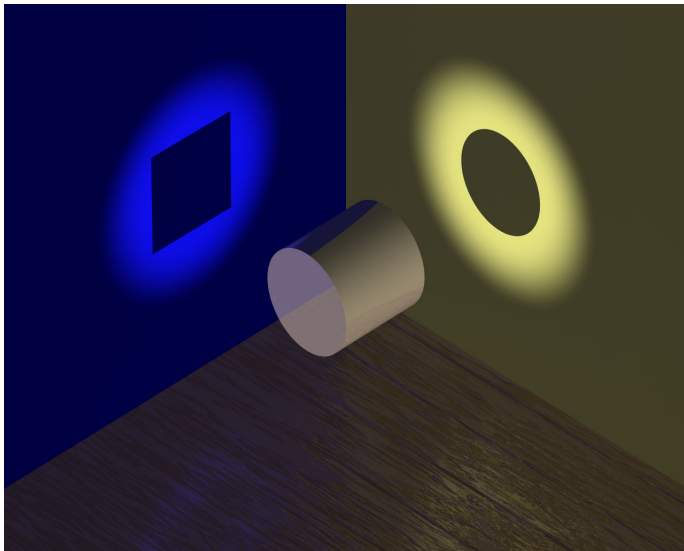
Image d'un nuage de points dans \mathbb{R}^3 : Vol d'oiseaux



Photo = représentation en **2D** d'une scène en **3D**

1. Blog de Tiguert : <http://taourirt-yakoub-guenzet.over-blog.fr/>

Exemple dans \mathbb{R}^3 : Représentation d'un cylindre



Exemple 1 : Représentation sur un plan (HLP²)

Le point de vue adopté par le photographe rend-il convenablement la **forme** du nuage dans l'espace ?

Husson, Lê, Pagès

Analyse de données avec R

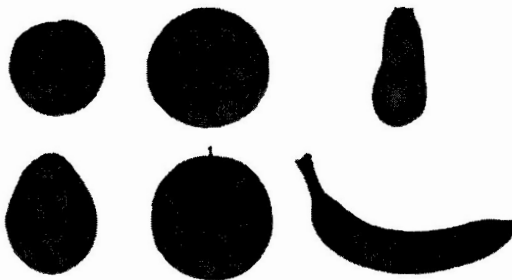


Exemple 1 : Représentation sur un plan (HLP³)

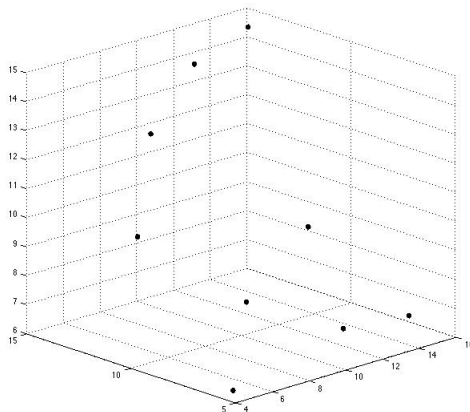
Le point de vue adopté par le photographe rend-il convenablement la **forme** du nuage dans l'espace ?

Husson, Lê, Pagès

Analyse de données avec R

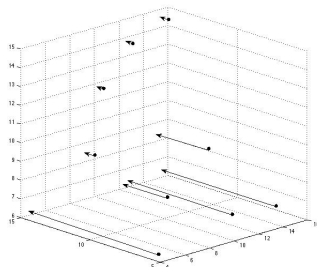
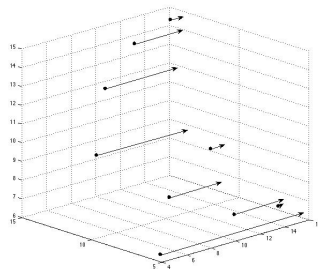
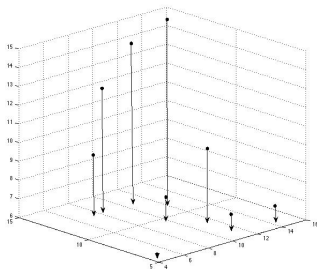


Exemple 2 : Nuage de points dans \mathbb{R}^3



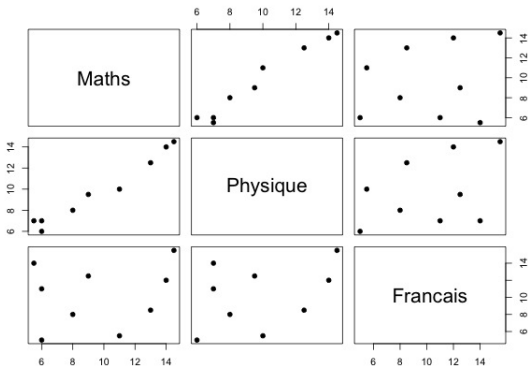
Forme du nuage difficilement visible.

Projections du nuage



Projections du nuage (2)

Représentation matricielle (*Scatterplot*) : projections sur les plans définis par les axes du repère de \mathbb{R}^3 pris 2 à 2.



Visualisation et interprétation **globales** du nuage difficiles.

Exemples dans \mathbb{R}^3 : en résumé

- **Etudier les individus** du point de vue de **l'ensemble des variables**.
- **Représentation sur un plan** :
 - Trouver le meilleur point de vue afin de rendre compte **au mieux** de la scène photographiée.
 - Déterminer le **meilleur plan** sur lequel **représenter** les points d'un espace de dimension 3, et qui rendra **au mieux** la **forme** du nuage de départ.
- **Généralisation** aux espaces de dimension $p > 3$: **projection** du nuage de points dans \mathbb{R}^p sur des sous-espaces de dimensions $q \ll p$.
- Sous-espaces de dimensions $q \ll p$ construits **sur les données** de manière à conserver **au mieux** la **forme** du nuage dans \mathbb{R}^p .

Objectifs de l'ACP

- **Représenter des données de grande taille**⁴ pour :
 - la visualisation → description résumée
 - des individus (détection d'individus ou groupes d'individus atypiques),
 - des variables (liaisons et sélection).
 - la compression → réduction du nombre de variables,
 - le débruitage → réduction de la variabilité.
- **Citation** (Lebart, Piron, Morineau, Chapitre 1)

"Nous cherchons en fait une technique de réduction s'appliquant de façon systématique à divers types de tableaux et conduisant à une reconstitution rapide mais approximative du tableau de départ."

4. sans hypothèse de type statistique, ou de modélisation

Outils

- Construction de **nouvelles variables (=facteurs)** concentrant la variance du nuage de points en un petit nombre q de facteurs.
- **Représentation graphique des individus** dans un sous-espace de dimension $q \ll p$ minimisant les déformations du nuage de points.
- **Représentation graphique des variables** dans un sous-espace de faible dimension $q \ll p$ explicitant les liaisons initiales entre ces variables.
- **Réduction de la dimension (=compression)** : **approximation** du tableau de données initial $n \times p$ par un tableau $n \times q$, avec $q \ll p$.

Principe et Définitions

- Effectuer un **changement de repère** dans \mathbb{R}^p (individus), ou \mathbb{R}^n (variables), de manière à **concentrer la variabilité** du nuage de points sur les premiers **axes factoriels** du nouveau repère.
- Construire les **nouvelles variables** sur les **axes factoriels**.
- **Facteurs principaux** $(F^{(k)})_{1 \leq k \leq q}$:

- **combinaisons linéaires** des variables initiales : pour $k = 1, \dots, q$

$$\begin{aligned} F^{(k)} &= a_{0,k} + a_{1,k}X^{(1)} + a_{2,k}X^{(2)} + \dots + a_{p,k}X^{(p)} \\ &= a_{0,k} + \sum_{j=1}^p a_{j,k}X^{(j)} \end{aligned}$$

- **2 à 2 non-corrélés** : $\forall k \neq k' \quad \text{Cor}(F_k, F_{k'}) = 0$
- **Composantes principales** $(c^{(1)}, \dots, c^{(q)})$: mesures des individus sur les nouvelles variables $(F^{(1)}, \dots, F^{(q)})$

$$\forall i \in \{1, \dots, n\} \quad \forall k \in \{1, \dots, q\} \quad c_i^{(k)} = a_{k,0} + \sum_{j=1}^p a_{k,j}x_i^{(j)}$$

Construction des facteurs principaux

- 1^{ère} **composante principale** :
 - Combinaison linéaire des variables expliquant le mieux la variabilité de l'échantillon.
 - Déterminée par la direction dans laquelle le nuage de points a son allongement maximum.
- 2^{ème} **composante principale** :
 - Combinaison linéaire des variables expliquant le mieux la variance résiduelle.
 - Orthogonale à la précédente.
- Etc...

Notations et définitions

- $X = (x_i^{(j)})_{1 \leq i \leq n, 1 \leq j \leq p}$ matrice $n \times p$ (n individus et p variables).
- g **centre de gravité** des n points de \mathbb{R}^p ,

$$g = (\bar{x}^{(1)}, \dots, \bar{x}^{(p)})$$

- $\langle \cdot, \cdot \rangle$ **produit scalaire** sur \mathbb{R}^p ,

$$\langle X_A, X_B \rangle = \langle \overrightarrow{OX_A}, \overrightarrow{OX_B} \rangle = \sum_{j=1}^p x_A^{(j)} x_B^{(j)}$$

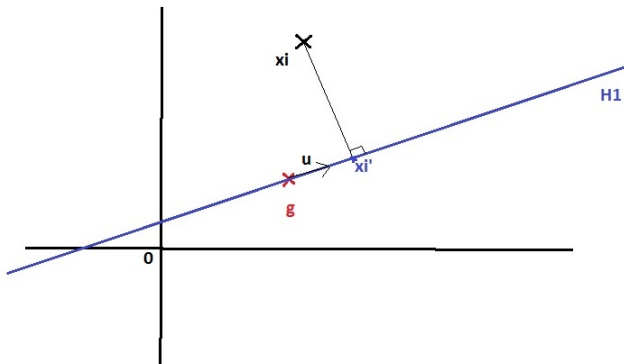
- d **distance euclidienne** sur \mathbb{R}^p ,

$$\begin{aligned} d^2(X_A, X_B) &= \langle X_A - X_B, X_A - X_B \rangle = \|X_A - X_B\|^2 \\ &= \sum_{j=1}^p \left(x_A^{(j)} - x_B^{(j)} \right)^2 \end{aligned}$$

Notations et définitions (2)

- H_1 droite de vecteur directeur unitaire \vec{u} , **passant par g** .
 x'_i **projection orthogonale** de x_i sur H_1 est définie par

$$\begin{aligned} \langle \overrightarrow{x_i x'_i}, \vec{u} \rangle &= 0 \\ |\langle \overrightarrow{g x'_i}, \vec{u} \rangle| &= d(g, x'_i) \end{aligned}$$



Première composante principale : principe général

- H_1 sous-espace de dimension 1 (= droite) de \mathbb{R}^p **passant par g** .
 $I_{exp}(H_1)$ **inertie expliquée** par H_1

$$I_{exp}(H_1) = \frac{1}{n} \sum_{i=1}^n d^2(g, x'_i)$$

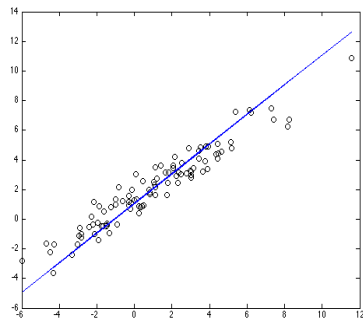
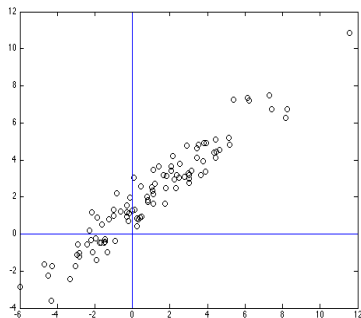
où x'_i projection orthogonale de x_i sur H_1 .

- **Maximiser l'inertie expliquée** sur l'ensemble des droites H_1 passant par g .
- C_1 tel que

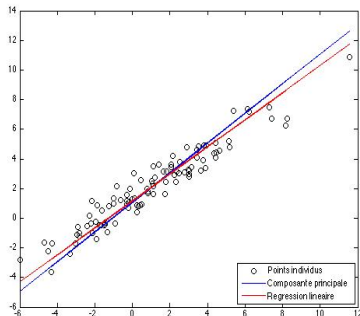
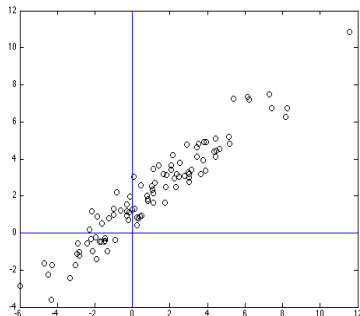
$$I_{exp}(C_1) = \max_{\{H_1 \text{ droite de } \mathbb{R}^p \text{ passant par } g\}} I_{exp}(H_1)$$

- $\implies C_1$ droite obtenue comme combinaison linéaire des variables maximisant la somme des distances à la nouvelle origine g .
- C_1 est la **"meilleure" droite** expliquant la variance du nuage de points (droite de variance maximale).

Première composante principale (exemple en 2D)



Première composante principale (exemple en 2D)



Attention ! De manière générale, lorsque $p = 2$, la droite de régression linéaire **n'est pas égale** au premier axe factoriel C_1 .

De 1 à q facteurs

- **Généralisation** à tout sous-espace H_q de dimension $q \leq p$ **contenant** g .
- u_1, \dots, u_q vecteurs unitaires orthogonaux engendrant H_q .
 x'_i **projection orthogonale** de x_i sur H_q est définie par

$$\forall k = 1, \dots, q \quad \langle \overrightarrow{x_i x'_i}, \vec{u}_k \rangle = 0$$

- **Inertie expliquée** du sous-espace H_q

$$I_{exp}(H_q) = \frac{1}{n} \sum_{i=1}^n d^2(g, x'_i)$$

- **Inertie résiduelle** du sous-espace H_q

$$I_{res}(H_q) = \frac{1}{n} \sum_{i=1}^n d^2(x_i, x'_i)$$

Décomposition de l'inertie

- **Sous-espace factoriel** (C_1, \dots, C_q) de dimension q : sous-espace E_q de dimension q tel que

$$I_{exp}(E_q) = \max_{\{H_q \text{ sous-espace de dimension } q \text{ d'origine } g\}} I_{exp}(H_q)$$

- **Inertie résiduelle** de E_q

$$I_{res}(E_q) = \frac{1}{n} \sum_{i=1}^n d^2(x_i, x'_i)$$

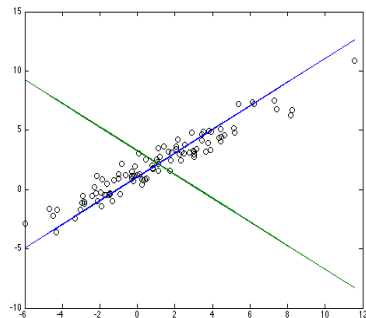
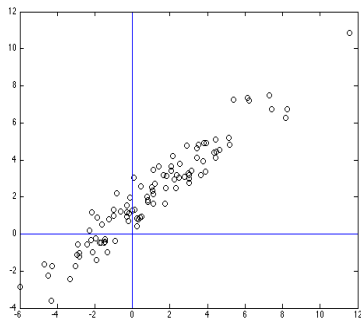
où x'_i projection orthogonale de x_i sur E_q .

- **Inertie totale** du nuage de points $I_{tot} = \sum_{j=1}^p \text{Var}(X^{(j)})$
- **Théorème de Pythagore** : $I_{tot} = I_{exp}(E_q) + I_{res}(E_q)$

Construction du repère factoriel : Principe

- **Changement de repère** : construction d'un nouveau repère de \mathbb{R}^p
 - d'**origine** $g = (\bar{x}^{(1)}, \dots, \bar{x}^{(p)})$,
 - **concentrant la variance** du nuage de points sur les premiers axes,
 - **minimisant les déformations** du nuage de points sur les premiers axes.
- Construction basée sur la **diagonalisation** de la matrice des variances/covariances (ou des corrélations)
 - Part de variance expliquée par chaque facteur donnée par les **valeurs propres** de la matrice des variances/covariances (ou des corrélations).
 - Nouveau repère défini par les **vecteurs propres** de la matrice des variances/covariances (ou des corrélations).

Changement de repère (exemple en 2D)



Construction du repère factoriel : définitions

Soit M matrice carrée d'ordre p .

- $\lambda \neq 0$ est une **valeur propre** de M s'il existe un vecteur **non-nul** $u = (u_1, \dots, u_p)$ tel que

$$Mu^t = \lambda u^t.$$

- Un tel vecteur u est un **vecteur propre** de M pour la valeur propre λ .
- Le **sous-espace propre** E_λ associé à la valeur propre λ est le sous-espace de \mathbb{R}^p engendré par l'ensemble des vecteurs propres de M pour la valeur propre λ . La **dimension** de E_λ est appelée **multiplicité** de la valeur propre λ .
- M est dite **diagonalisable** si la somme de ses sous-espaces propres est égale à \mathbb{R}^p :

$$E_{\lambda_1} + E_{\lambda_2} + \dots + E_{\lambda_r} = \mathbb{R}^p$$

Dans ce cas, il existe une **base** de \mathbb{R}^p constituée de **vecteurs propres** de M .

Résultats mathématiques

V matrice des **variances/covariances** et Σ des **corrélations** de X :

$$\begin{aligned}\forall (j, j') \in \{1; \dots; p\}^2 \quad V_{j,j'} &= \text{Cov}(X^{(j)}, X^{(j')}) \quad (\text{covariance}) \\ \Sigma_{j,j'} &= \text{Cor}(X^{(j)}, X^{(j')}) \quad (\text{corrélation})\end{aligned}$$

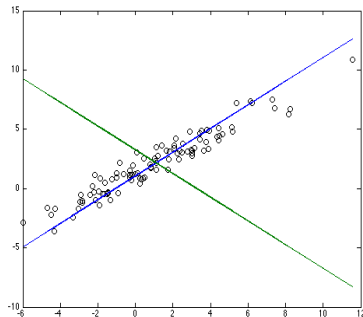
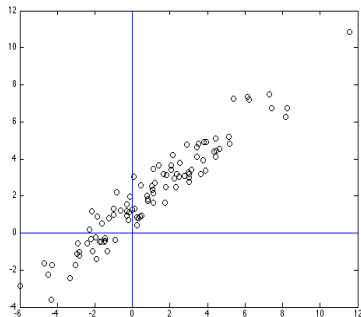
On montre que

- V et Σ sont diagonalisables, de valeurs propres réelles et positives.
- **les axes factoriels** $C^{(1)}, \dots, C^{(q)}$ sont engendrés par les q vecteurs propres **unitaires** u_1, \dots, u_q associés aux q plus grandes valeurs propres $\lambda_1 \geq \dots \geq \lambda_q$ de V (ou Σ) (multiplicité comprise).
- **l'inertie expliquée** par le sous-espace E_q est la somme des q plus grandes valeurs propres de V (ou Σ) : $I_{exp}(E_q) = \lambda_1 + \dots + \lambda_q$.
- $E_p = \mathbb{R}^p$ ($q=p$) : changement de repère de manière à faire porter la variance du nuage de points sur les premiers axes.

Résultats mathématiques : Résumé

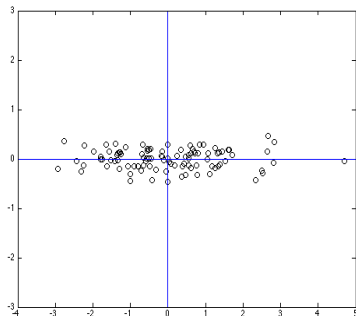
- Pour $1 \leq q \leq p$, E_q sous-espace engendré par les q premiers axes factoriels, de repère $(g, C^{(1)}, \dots, C^{(q)})$.
- E_q sous-espace de dimension q **concentrant le plus de variabilité** du nuage de points.
- **L'inertie expliquée** de E_q vérifie $I_{exp}(E_q) = \lambda_1 + \lambda_2 + \dots + \lambda_q$.
- $I_{exp}(E_q)$ augmente avec q , alors que l'inertie résiduelle $I_{res}(E_q)$ diminue avec q .
- On en déduit donc : $I_{tot} = \dots\dots\dots$

Changement de repère (exemple en 2D)

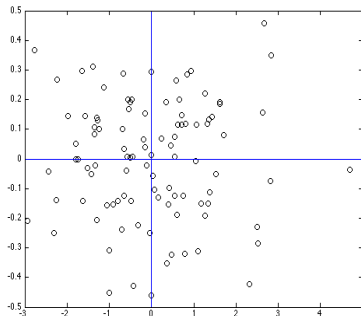


- **Axe 1** : associé à la plus grande valeur propre, portant la plus grande part de variance du nuage de points.
- **Axe 2** : associé à la plus petite valeur propre, portant la part de variance résiduelle du nuage de points.

Changement d'échelle (exemple en 2D)



Rotation du repère...



... et changement d'échelle.

Facteurs et Composantes Principaux

- **Facteurs principaux** $(F^{(1)}, \dots, F^{(p)})$: nouvelles variables définies par les axes factoriels, combinaisons linéaires des variables d'origine $(X^{(1)}, \dots, X^{(p)})$. Pour $k = 1, \dots, p$,
 - $\overline{F^{(k)}} = 0$,
 - $\text{Var}(F^{(k)}) = \lambda_k$,
 - Pour $k' \neq k$, $\text{Cor}(F^{(k)}, F^{(k')}) = 0$.

- **Composantes principales** $(c^{(1)}, \dots, c^{(p)})$: mesures des individus sur les nouvelles variables $(F^{(1)}, \dots, F^{(p)})$.

Soit $Y = (y_1, \dots, y_n)$ la table de données

- centrées si on utilise V la matrice des variances/covariances
- centrées et réduites si on utilise Σ la matrice des corrélations

Pour $i = 1, \dots, n$ et $k = 1, \dots, p$

$$c_i^{(k)} = \langle y_i, u_k \rangle$$

- On obtient un **tableau transformé** $(c_i^{(k)})_{1 \leq i \leq n, 1 \leq k \leq p}$ à partir du tableau d'origine $X = (x_i^{(j)})_{1 \leq i \leq n, 1 \leq j \leq p}$.

Axes factoriels : interprétations

- **Part d'inertie/variance expliquée** par le $k^{\text{ème}}$ axe factoriel :

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

- **Part d'inertie/variance expliquée** par les q premiers axes factoriels :

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

- **Remarque** : Quand on utilise Σ la matrice des **corrélations** de X , alors

$$\sum_{k=1}^p \lambda_k = \text{Tr}(\Sigma) = \dots\dots\dots$$

Notion de Contribution

- **Contribution** de l'individu i à l'axe k :

$$100 \frac{\left(c_i^{(k)}\right)^2}{n \times \lambda_k}$$

La contribution permet de déceler quels individus interviennent dans l'élaboration d'un axe.

Généralement, on analyse les contributions dont la valeur est supérieure à la moyenne, à savoir $\frac{1}{n}$.

Plus la contribution de l'individu i sur l'axe k est élevée, plus cet individu intervient dans la création du facteur k .

Notion de Qualité de représentation

- **Qualité de représentation** de l'individu i sur le sous-espace E_q :

$$Q_i^q = \frac{\sum_{k=1}^q \left(c_i^{(k)}\right)^2}{\sum_{k=1}^p \left(c_i^{(k)}\right)^2}$$

La qualité de la représentation permet de déceler sur quels axes un individu est bien représenté.


La qualité de représentation sur le plan des axes (j,k) est donné par :

$$Q_i^{j,k} = Q_i^j + Q_i^k$$

On évite d'interpréter les individus ayant une faible qualité de représentation sur un plan.

Exemple

- Les données (d'après P. Besse et A. Baccini) : notes (de 0 à 20) obtenues par 9 élèves dans 4 disciplines.

 **VIEWTABLE(Nouveau): (Enregistrer sous: Work.No**

	Nom	Maths	Phys	Fran	Angl
1	Olivier	6	6	5	5.5
2	Eva	8	8	8	8
3	Mael	6	7	11	9.5
4	Djamel	14.5	14.5	15.5	15
5	Lise	14	14	12	12.5
6	Farah	11	10	5.5	7
7	Swann	5.5	7	14	11.5
8	Salma	13	12.5	8.5	9.5
9	Djibril	9	9.5	12.5	12

- Tableau à 9 lignes et 4 colonnes
- Matrice X de taille 9×4

Exemple : Résultats

- V = Matrice des variances-covariances de X :

$$\begin{pmatrix} 12,81 & 11,16 & 2,99 & 5,43 \\ 11,16 & 10,06 & 4,64 & 6,17 \\ 2,99 & 4,64 & 13,57 & 10,45 \\ 5,43 & 6,17 & 10,45 & 8,9 \end{pmatrix}$$

- Σ = Matrice des corrélations de X :

$$\begin{pmatrix} 1 & 0,98 & 0,23 & 0,5 \\ 0,98 & 1 & 0,4 & 0,65 \\ 0,23 & 0,4 & 1 & 0,95 \\ 0,5 & 0,65 & 0,95 & 1 \end{pmatrix}$$

Exemple : Résultats (2)

- **ACP réalisée sur la matrice des corrélations**
- Valeurs propres obtenues :
 $2,876 ; 1,12 ; 0,003 ; 0,001$
- Somme des variabilités sur facteurs :
 $2,876 + 1,12 + 0,003 + 0,001 = 4$
- On retrouve bien que la somme des valeurs propres vaut 4, le nombre de variables.

Exemple : Lecture des valeurs propres

- Pourcentages de variance expliquée :

Valeur propre	% variance	% variance cumulé
2,876	71,89	71,89
1,12	27,99	99,88
0,003	0,09	99,97
0,001	0,03	100

- ⇒ les 2 premiers facteurs expliquent 99,88% de la variance totale.

Exemple : Contribution et qualité de représentation

Contributions et qualité de représentation des individus :

	Dim.1	Dim.2	Dim.3	Dim.4		Dim.1	Dim.2	Dim.3	Dim.4
Olivier	29.0678526	1.8126836	1.64709836	5.4220465	Olivier	0.9761627	0.02370208	6.881416e-05	6.637524e-05
Eva	5.9470877	0.2315680	0.05976426	5.2517423	Eva	0.9847410	0.01492971	1.231142e-05	3.169966e-04
Mael	4.1062383	10.9257210	10.55245441	0.1324655	Mael	0.4903855	0.50804096	1.567819e-03	5.766733e-06
Djamel	38.0502446	0.3419015	0.40431127	23.7872876	Djamel	0.9962742	0.00348560	1.317002e-05	2.270387e-04
Lise	16.2609747	3.9113513	1.87097495	37.9695950	Lise	0.9135338	0.08555789	1.307660e-04	7.775845e-04
Farah	3.6425278	22.2536569	2.12644014	19.2458911	Farah	0.2957329	0.70348268	2.147825e-04	5.695980e-04
Swann	0.4329506	37.2483418	9.46360493	0.9073247	Swann	0.0289634	0.97022686	7.876200e-04	2.212623e-05
Salma	1.4860993	16.5362345	13.62632159	1.6275229	Salma	0.1871131	0.81067773	2.134441e-03	7.469940e-05
Djibril	1.0060244	6.7385415	60.24903009	5.6561246	Djibril	0.2714009	0.70782185	2.022097e-02	5.562313e-04

Covariance ou Corrélation ?

- **Corrélation**

- revient à travailler avec les variables centrées et réduites
- les variables sont sans dimension
- les variables ont toutes la même dispersion
- on réalise ainsi une **ACP normée**

- **Covariance**

- revient à travailler avec les variables centrées
- les variables sont directement analysées
- les variables ont des dispersions et des dimensions différentes
- à n'utiliser que si les données sont homogènes
- on réalise ainsi une **ACP non-normée**

Visualisation des données

- **Premier plan factoriel**

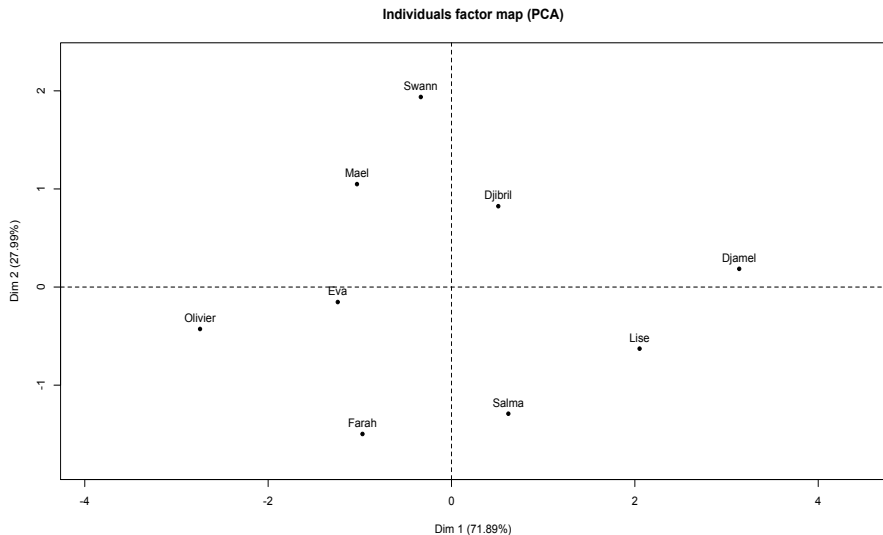
- **Meilleur plan** au sens des **moindres carrés**.
- Plan engendré par les **2 premiers axes factoriels** $C^{(1)}$ et $C^{(2)}$.
- Plan associé aux **2 plus grandes valeurs propres** λ_1 et λ_2 de la matrice Σ .

- **Autres plans factoriels** : plans engendrés par les premiers axes $C^{(1)}$ et $C^{(k)}$, $2 \leq k \leq q$, avec q convenablement choisi en terme de variance expliquée par les q premiers facteurs.

- **Visualisation simple** permettant de se rendre compte de la forme du nuage dans des espaces de dimension 2.

- Remarque : On représente toujours le facteur 1 en abscisses (se méfier notamment des étiquettes sous R)

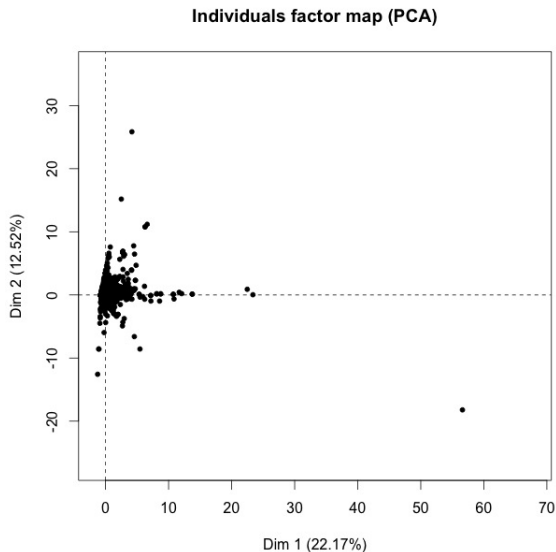
Exemple : représentation dans le premier plan factoriel



Individus atypiques

- Les **individus atypiques** sont repérés sur les axes factoriels par des **contributions très élevées** par rapport à celles des autres individus.
- Ces individus ont tendance à **écraser** les autres de part ces contributions extrêmes.
- **Que faire ?**
 - repérer les individus atypiques, i.e. ceux ayant des contributions trop importantes par rapport à celles des autres.
 - si le nombre d'individus atypiques représente un faible pourcentage de l'échantillon, les éliminer de l'étude.
 - si ce nombre est élevé, les traiter comme un groupe à part entière.

Individus atypiques : exemple sur le premier plan factoriel



Qualité de représentation du premier plan factoriel

- **Premier plan factoriel** : plan concentrant la plus grande part de variance du nuage de points.
- **Toujours garder en tête** la part de variance expliquée par le premier plan factoriel avant toute **interprétation hâtive**.
- Si la part de variance expliquée par le premier plan factoriel est faible, regarder la répartition des individus sur d'autres plans concentrant moins de variance.
- **Exemple des notes** : le premier plan factoriel explique 99,88% de la variance totale, sa qualité de représentation est très bonne.
- **Exemple des individus atypiques** : le premier plan factoriel n'explique que 34,68% de la variance totale, sa qualité de représentation est faible.

Point de vue des variables ($p \leq n$)

- variables regardées comme des points de \mathbb{R}^n
- $d(j, j')$, distance dans \mathbb{R}^n entre les variables $X^{(j)}$ et $X^{(j')}$.
- On diagonalise Σ^t , transposée de la matrice des **corrélations** de X .
- **On montre que** :
 - la **norme** des vecteurs variables dans \mathbb{R}^n est égale à **1**
 - $d^2(j, j') = 2 (1 - \text{Cor}(X^{(j)}, X^{(j')}))$
 - $\text{Cor}(X^{(j)}, X^{(j')}) = \cos(X^{(j)}, X^{(j')})$
 - les variables sont décrites sur la sphère unité de $\mathbb{R}^p \subset \mathbb{R}^n$ par leurs contributions sur les facteurs-individus
- **Ce qui implique** :
 - les vecteurs variables sont tous sur la sphère unité de \mathbb{R}^p
 - plus les variables sont corrélées positivement, et plus elles sont proches
 - moins les variables sont corrélées et plus les vecteurs variables sont orthogonaux

Sphère et cercle des corrélations (LPM⁵)

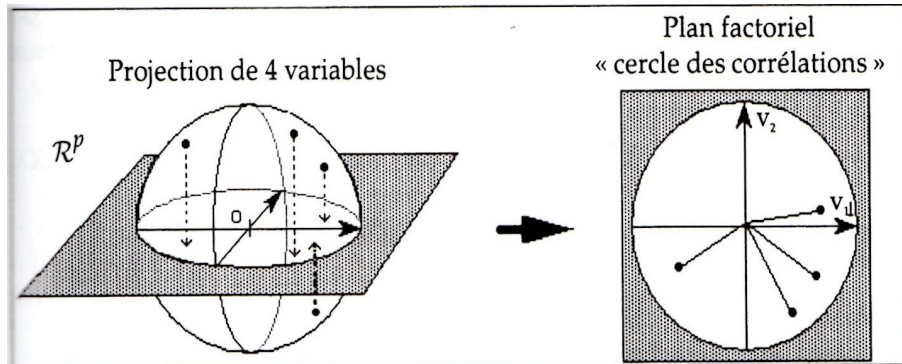


Figure 3.2 – 5. Représentation de la sphère et du cercle des corrélations

Contribution des variables aux axes factoriels

- **Coordonnée** de la variable $X^{(j)}$ sur l'axe $C^{(k)}$

$$b_{j,k} = \text{Cor}(F^{(k)}, X^{(j)}) = \frac{\sqrt{\lambda_k}}{\sigma(X^{(j)})} u_k^{(j)}$$

- **Contribution de la variable j à l'axe k :**

$$100 \frac{(b_{j,k})^2}{\lambda_k}$$

Plus sa valeur est élevée, et plus la variable contribue à la construction de l'axe.

- De plus, on a :

$$\cos(C^{(k)}, X^{(j)}) = \text{Cor}(F^{(k)}, X^{(j)}) = b_{j,k}$$

- \implies contribution de la variable j à l'axe k interprétée en terme de **proximité** de la variable à l'axe en question.
- **Attention** : si la contribution d'une variable est faible, cela signifie qu'elle contribue à d'autres axes.

Cercle des corrélations : interprétation

- une variable est **bien représentée** sur le cercle des corrélations si son vecteur est proche du cercle.
- Seules les variables bien représentées interviennent dans l'interprétation des axes.
- Une variable contribue **positivement** à l'axe si $b_{j,k} \geq 0$, **négativement** sinon.
- Les variables **orthogonales** sont non-corrélées.
- Deux variables sont **corrélées** si leurs vecteurs variables sont approximativement colinéaires.

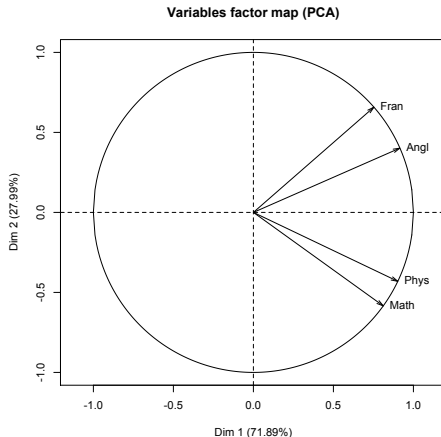
Exemple : contributions

- Coordonnées des variables dans le repère factoriel :

	u_1	u_2	u_3	u_4
Math	0,81	-0,58	-0,01	0,02
Phys	0,9	-0,43	0,03	-0,02
Fran	0,75	0,66	0,03	0,01
Angl	0,91	0,4	-0,04	-0,01

⇒ les variables sont portées par les 2 premiers axes factoriels.

Exemple : cercle des corrélations



- Axe 1 : Bon/mauvais
- Axe 2 : Lettres/Sciences

Représentation simultanée individus/variables

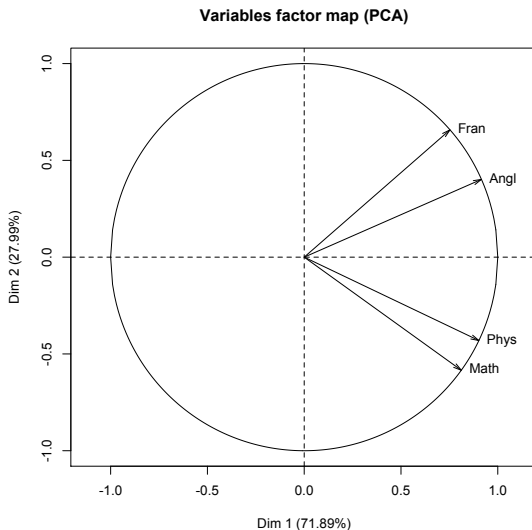
- **Représentation simultanée**

- des individus
- des variables

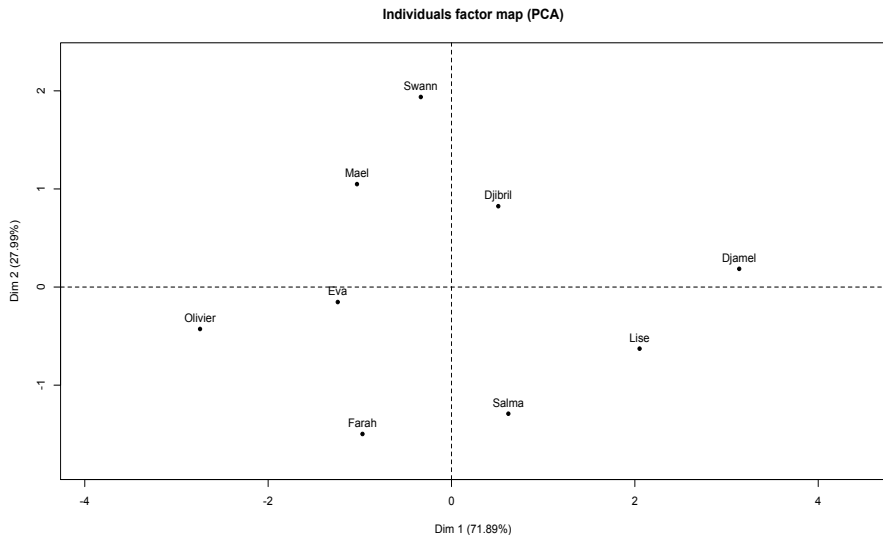
dans les premiers plans factoriels.

- Permet de donner une **signification** aux axes (variables).
- Permet de visualiser les **contributions** de chaque individu sur chaque axe (individus).
- Permet de **décrire** les données en prenant en compte **l'ensemble des p variables** en croisant la représentation des individus avec celle des variables.
- **Attention !** Individus et variables vivent dans des espaces **différents**. Il est donc **incorrect** de les représenter sur le même graphique.

Exemple : cercle des corrélations...



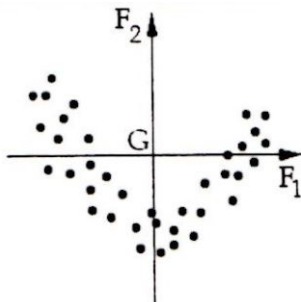
... avec la représentation dans le premier plan factoriel



Effet Guttman

- S'il existe une ou plusieurs variables dans le tableau de données **fortement corrélées à toutes les autres**, on parle d'**effet Guttman**.
- Effet **immédiatement visible** sur le premier plan factoriel : les points sont distribués suivant une parabole.
- Effet biaisant l'analyse des données multidimensionnelles, et aboutissant à une **analyse erronée**.
- Dans ce cas, **isoler** la ou les variables fautives, et les **retirer** de l'étude.

Effet Guttman : Exemple



- L'axe 1 oppose les extrêmes entre eux.
- L'axe 2 oppose les extrêmes aux moyens.

Réduction de la dimension : choix du nombre de facteurs

- **Pas de "recette"** universelle pour choisir le nombre de facteurs à retenir.
- **Choix dépendant de l'analyse :**
 - Garder un petit nombre de facteurs concentrant l'essentiel de la variabilité pour représenter les données
 - Préparer les données pour une éventuelle analyse ultérieure. Dans ce cas, ce n'est pas un inconvénient de garder beaucoup de facteurs.

Choix du nombre de facteurs : méthodes numériques

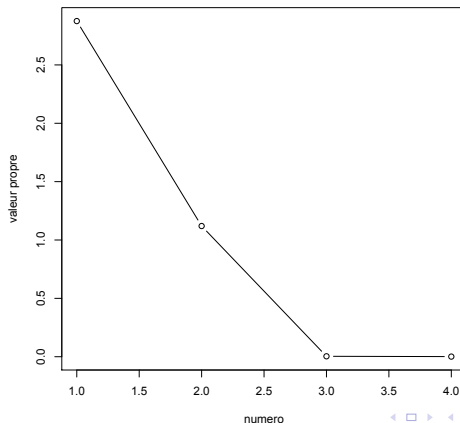
1) Règle de Kaiser : garder les axes correspondant aux valeurs propres supérieures à la moyenne des valeurs propres, soit

- supérieures à dans le cas de la matrice des covariances,
- supérieures à dans le cas de la matrice des corrélations.

2) Règle empirique : garder les facteurs expliquant un pourcentage de variance cumulé satisfaisant (généralement 80%).

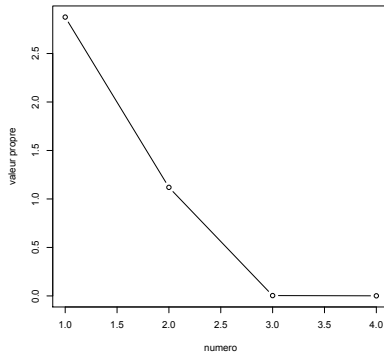
Choix du nombre de facteurs : méthode graphique

3) **Règle de l'ébouli des valeurs propres** : garder les axes correspondant aux valeurs propres situées **avant le point d'inflexion** sur le graphe des valeurs propres.



Exemple : choix du nombre de facteurs

- Valeurs propres de la matrice des corrélations :
2,876 ; 1,12 ; 0,003 ; 0,001
- Ebouli des valeurs propres :



- \Rightarrow Il semble judicieux de garder facteurs.

Variables illustratives

- Les variables illustratives n'interviennent pas dans la création des axes. Mais on peut les représenter graphiquement.
- **Variables quantitatives**
 - Si on ne souhaite pas voir des variables influencer sur la création des axes
 - Elles sont représentées sur le cercle des corrélations.
- **Variables qualitatives**
 - Chacune des modalités est représentée dans le plan factoriel des individus
 - Une modalité est localisée au barycentre des individus possédant cette modalité, et représente un individu moyen.
- Des individus peuvent également être traités comme illustratifs.

ACP : Etude de cas

Obs country	viandr	viandb	oeuf	lait	poisson	cereals	feculent	oleagine	fruitleg
1 Bulgaria	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2
2 Yugoslavia	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2
3 Romania	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8
4 Russian Federat	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9
5 Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
6 Greece	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5
7 Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2
8 Italy	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
9 Poland	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6
10 Czech Republic	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0
11 Spain	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2
12 France	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
13 Liechtenstein	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3
14 Portugal	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9
15 Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0
16 Iceland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4
17 Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
18 Ukraine	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
19 United Kingdom	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
20 Ireland	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9
21 Norway	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7
22 Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
23 Denmark	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4
24 Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0
25 Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8

- Pour chaque pays
- Les consommations des 9 protéines.

ACP : Etude de cas

Questions

- ❶ Quelle est la part de variance expliquée par le premier plan factoriel ?
- ❷ Combien d'axes est-il pertinent de conserver :
 - via la règle de Kaiser
 - via l'ébouli des valeurs propres ?
- ❸ Quels pays contribuent le plus à la construction de l'axe 1 ? L'axe 2 ?
- ❹ Quels sont les pays les mieux représentés sur le premier plan factoriel ?
Et les moins bien représentés ?
- ❺ Interpréter l'axe 1 en termes de variables et d'individus.
Faire de même pour les autres axes conservés.

Exemple sous R : Commandes et sorties

- **Commandes :**

- `library(FactoMineR)`
charge le package *FactoMineR*.
- `proteine.acp = PCA(proteine[,-1], ncp = 9, graph = T)`
effectue l'ACP sur les données avec 9 (= nombre de variables) facteurs principaux (défaut = 5). Trace le graphe des données dans le premier plan factoriel, ainsi que le cercle des corrélations (défaut = T).

- **Sorties principales :**

- `proteine.acp$eig` : tableau des valeurs propres et des pourcentages de variance expliquée pour chaque facteur.
- `proteine.acp$var` : liste des coordonnées, contributions et statistiques des variables sur chaque facteur.
- `proteine.acp$ind` : liste des coordonnées, contributions et statistiques des individus sur chaque facteur.

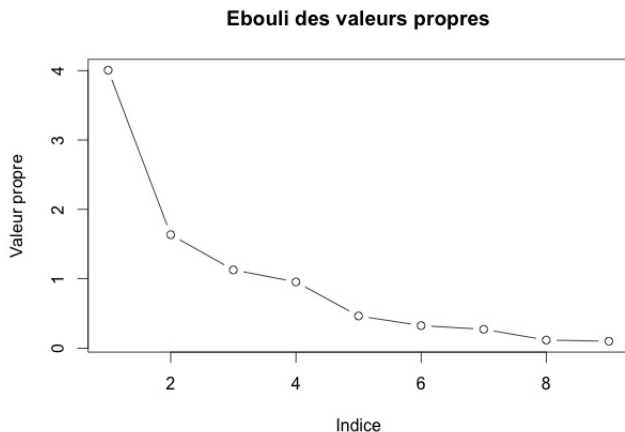
Exemple sous R : Sorties de la fonction PCA

Contribution à la variance : `proteine.acp$eig`

```
> proteine.acp$eig
```

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.0064376	44.515973	44.51597
comp 2	1.6349994	18.166661	62.68263
comp 3	1.1279195	12.532439	75.21507
comp 4	0.9546640	10.607377	85.82245
comp 5	0.4638384	5.153760	90.97621
comp 6	0.3251310	3.612566	94.58878
comp 7	0.2716063	3.017848	97.60662
comp 8	0.1162919	1.292132	98.89876
comp 9	0.0991119	1.101243	100.00000

Ebouli des valeurs propres



Calcul des contributions

```
> proteines.acp$ind$contrib
```

	Dim.1	Dim.2	Dim.3	Dim.4
Bulgaria	10.21529473	4.31389487	8.455292e-02	0.20023233
Yugoslavia	13.65113880	2.74593330	1.568355e-01	2.94582788
Romania	7.90396117	3.18980912	1.814301e-02	1.65083779
Russian Federation	0.63695127	0.03126919	5.048478e-01	3.75521074
Albania	12.63361076	6.77488643	1.145894e+01	0.23019793
Greece	5.21687063	2.55382061	2.877676e+00	14.05199614
Hungary	2.20908090	1.69667813	1.353538e+01	0.20625790
Italy	2.44906973	0.40568488	5.873126e-02	6.52240726
Poland	0.01544616	0.72056470	8.034717e+00	0.91641205
Czech Republic	0.14273283	0.92560995	5.283610e+00	0.93959478
Spain	1.78966010	16.61694150	9.808519e-01	0.56314641
France	2.30263963	1.57186203	1.309276e-05	16.72333737
Liechtenstein	2.10493338	2.76290260	6.611433e+00	0.12332748
Portugal	3.02631468	46.87808620	7.032937e-03	3.48485950
Belgium	2.73620983	0.06482887	1.732067e-01	1.18346882
Iceland	2.54345520	0.90524760	1.551707e+01	8.74259615
Switzerland	0.86561150	1.43753641	8.789902e-02	5.97915787
Ukraine	2.10357670	0.51674548	6.267479e+00	5.63203480
United Kingdom	3.13194150	0.02250800	4.908279e+00	13.11843948
Ireland	7.37782674	1.48636032	1.460068e-03	0.82484732
Norway	0.98804087	1.72204731	1.072726e+01	5.64850989
Netherlands	2.80209770	2.11959139	2.170309e+00	0.06945830
Denmark	5.81823675	0.20764359	2.090502e+00	4.08411163
Sweden	2.77581758	0.10960328	6.055959e+00	2.35206994
Germany	4.55948085	0.21994428	2.387814e+00	0.05166026

Calcul des qualités de représentation

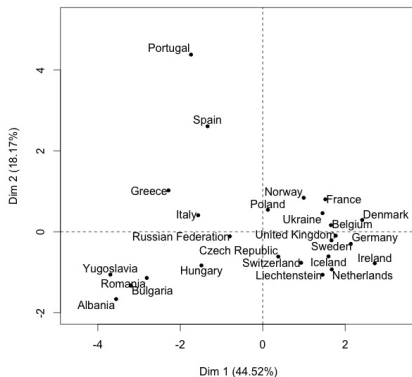
```
> proteines.acp$ind$cos2
```

	Dim.1	Dim.2	Dim.3	Dim.4
Bulgaria	0.74064685	0.127640667	1.725872e-03	0.003459287
Yugoslavia	0.86059920	0.070644974	2.783531e-03	0.044251865
Romania	0.80125581	0.131962435	5.177922e-04	0.039877043
Russian Federation	0.12801416	0.002564647	2.856485e-02	0.179836679
Albania	0.61215361	0.133965871	1.563137e-01	0.002657826
Greece	0.42667738	0.085239161	6.625995e-02	0.273854399
Hungary	0.27611042	0.086542586	4.762790e-01	0.006142907
Italy	0.44590063	0.030142896	3.010412e-03	0.282967784
Poland	0.00315723	0.060106008	4.623555e-01	0.044634288
Czech Republic	0.04557900	0.120622457	4.749976e-01	0.071494671
Spain	0.17226841	0.652747776	2.658020e-02	0.012916610
France	0.25375856	0.070691694	4.062055e-07	0.439146914
Liechtenstein	0.33799259	0.181047954	2.988712e-01	0.004718683
Portugal	0.12733142	0.804916064	8.330632e-05	0.034938070
Belgium	0.70929275	0.006858099	1.264039e-02	0.073101299
Iceland	0.23749029	0.034494383	4.078976e-01	0.194515555
Switzerland	0.20623455	0.139770809	5.895792e-03	0.339445900
Ukraine	0.32552406	0.032633275	2.730470e-01	0.207674046
United Kingdom	0.33322901	0.000977294	1.470206e-01	0.332585638
Ireland	0.77649844	0.063840380	4.326183e-05	0.020686063
Norway	0.15744344	0.111983485	4.812361e-01	0.214474524
Netherlands	0.53027881	0.163693855	1.156278e-01	0.003132112
Denmark	0.66284556	0.009653810	6.704883e-02	0.110869149
Sweden	0.45437109	0.007321542	2.790760e-01	0.091740721
Germany	0.79564030	0.015662940	1.173064e-01	0.002148078

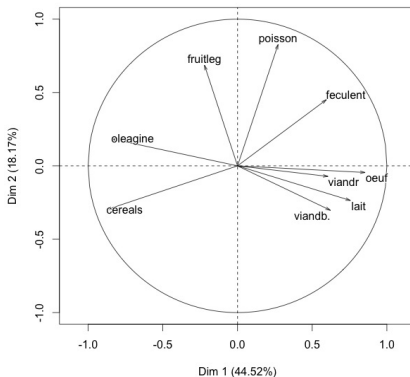
Représentation sur le premier plan factoriel

Représentation sur le premier plan factoriel et cercle des corrélations

Individuals factor map (PCA)



Variables factor map (PCA)

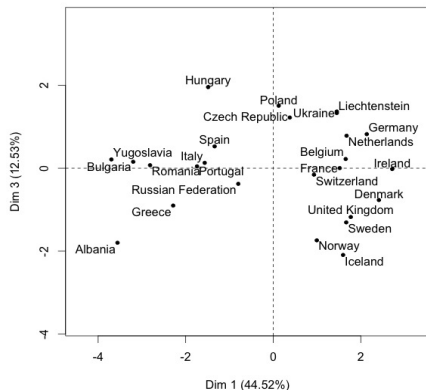


Représentation sur le deuxième plan factoriel

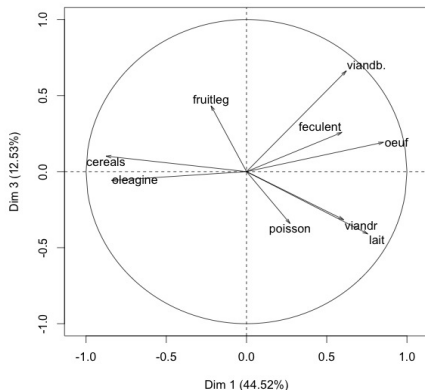
```
plot(proteine.acp, axes = c(1,3), choix ="ind")
```

```
plot(proteine.acp, axes = c(1,3), choix ="var")
```

Individuals factor map (PCA)



Variables factor map (PCA)

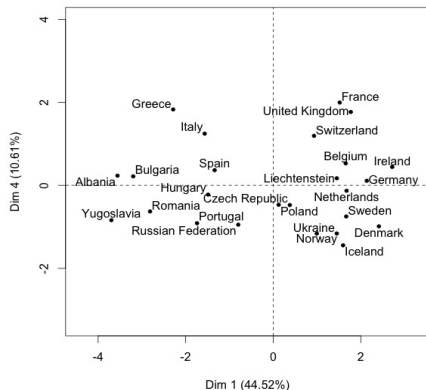


Représentation sur le troisième plan factoriel

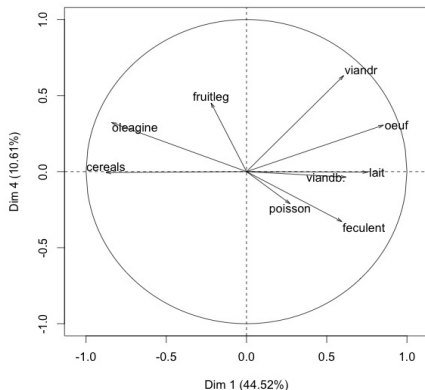
```
plot(proteine.acp, axes = c(1,4), choix ="ind")
```

```
plot(proteine.acp, axes = c(1,4), choix ="var")
```

Individuals factor map (PCA)



Variables factor map (PCA)



ACP : Exemple sous SAS

Procédure princomp :

```
/* *****  
/* 3. ACP sur données proteines */  
  
proc princomp data = proteines out = proteines_ACP;  
var viande -- fruitleg;  
run;
```

Table OUT : ajout à la table de données d'origine des coordonnées de chaque individu sur les facteurs principaux.

Options utiles :

- **vardef=n** : coefficients de la matrice Σ normalisés par n et non $n - 1$.
- **COV** : utilise la matrice de covariance au lieu de la matrice de corrélation.
- **STD** : standardise les coefficients sur les composants principaux dans la table OUT.
- **NOPRINT** : n'affiche pas les sorties.

Sortie de la procédure Princomp

The PRINCOMP Procedure

Observations	25
Variables	9

Simple Statistics

	viandr	viandb	oeuf	lait	poisson	cereals	feculent	oleagine	fruitleg
Mean	9.828000000	7.896000000	2.936000000	17.11200000	4.284000000	32.24800000	4.276000000	3.072000000	4.136000000
Std	3.347078328	3.694080851	1.117616511	7.10541577	3.402533370	10.97478625	1.634084861	1.985682083	1.803903176

Correlation Matrix

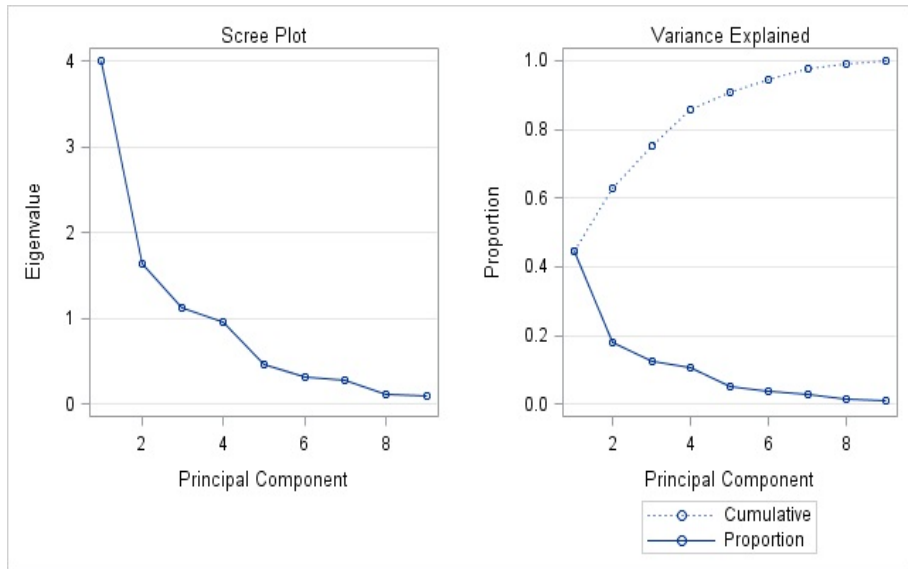
	viandr	viandb	oeuf	lait	poisson	cereals	feculent	oleagine	fruitleg
viandr	1.0000	0.1530	0.5856	0.5029	0.0610	-.4999	0.1354	-.3494	-.0742
viandb	0.1530	1.0000	0.6204	0.2815	-.2340	-.4138	0.3138	-.6350	-.0613
oeuf	0.5856	0.6204	1.0000	0.5755	0.0656	-.7124	0.4522	-.5598	-.0455
lait	0.5029	0.2815	0.5755	1.0000	0.1379	-.5927	0.2224	-.6211	-.4084
poisson	0.0610	-.2340	0.0656	0.1379	1.0000	-.5242	0.4039	-.1472	0.2661
cereals	-.4999	-.4138	-.7124	-.5927	-.5242	1.0000	-.5333	0.6510	0.0465
feculent	0.1354	0.3138	0.4522	0.2224	0.4039	-.5333	1.0000	-.4743	0.0844
oleagine	-.3494	-.6350	-.5598	-.6211	-.1472	0.6510	-.4743	1.0000	0.3750
fruitleg	-.0742	-.0613	-.0455	-.4084	0.2661	0.0465	0.0844	0.3750	1.0000

Sortie de la procédure Princomp (2)

Eigenvalues of the Correlation Matrix				
	Valeur propre	Différence	Proportion	Cumulée
1	4.00643757	2.37143813	0.4452	0.4452
2	1.63499945	0.50707994	0.1817	0.6268
3	1.12791950	0.17325554	0.1253	0.7522
4	0.95466396	0.49082557	0.1061	0.8582
5	0.46383840	0.13870742	0.0515	0.9098
6	0.32513097	0.05352464	0.0361	0.9459
7	0.27160633	0.15531443	0.0302	0.9761
8	0.11629190	0.01718000	0.0129	0.9890
9	0.09911190		0.0110	1.0000

Eigenvectors									
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9
viandr	0.302609	-0.056252	-0.297580	0.646477	0.322160	-0.459870	0.150334	0.019858	0.246000
viandb	0.310556	-0.236853	0.623897	-0.036992	-0.300165	-0.121007	-0.019664	0.027876	0.592397
oeuf	0.426679	-0.035336	0.181528	0.313164	0.079110	0.361249	-0.443272	0.491200	-0.333386
lait	0.377727	-0.184589	-0.385658	-0.003318	-0.200414	0.618438	0.462095	-0.081422	0.178084
poisson	0.135650	0.646820	-0.321274	-0.215955	-0.290031	-0.136791	-0.106394	0.448732	0.312826
cereals	-0.437743	-0.233485	0.095918	-0.006204	0.238168	0.080758	0.404964	0.702995	0.152260
feculent	0.297248	0.352826	0.242975	-0.336685	0.735973	0.147667	0.152753	-0.114540	0.121858
oleagine	-0.420334	0.143311	-0.054388	0.330288	0.150537	0.447010	-0.407262	-0.183800	0.518275
fruitleg	-0.110420	0.536190	0.407556	0.462056	-0.233517	0.118550	0.449978	-0.091963	-0.202950

Sortie de la procédure Princomp (3)



Représentation dans le premier plan factoriel

