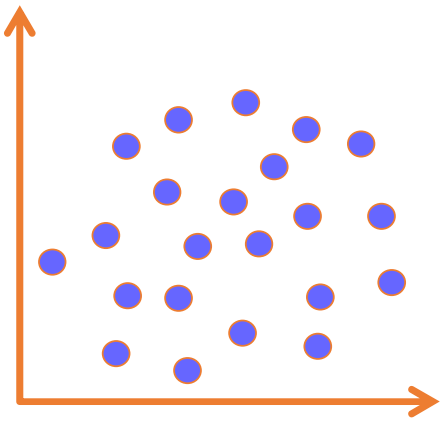
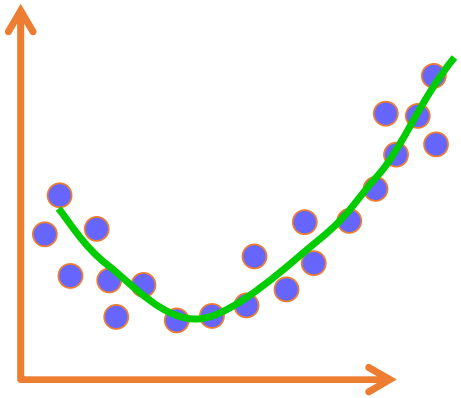
Analyses factorielles :
l'analyse en composantes principales

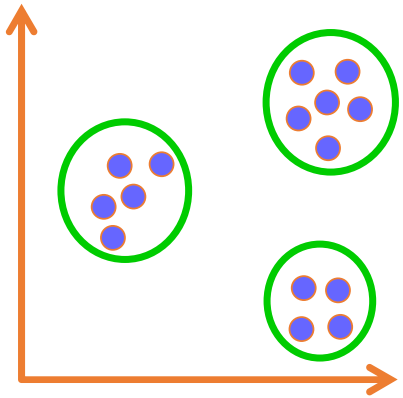
Marie-Camille CAUMON
Ingénieur de recherche
GeoRessources - UMR 7359
Entrée 3B - bureau A508
+33 3 72 74 55 37
marie-camille.caumon@univ-lorraine.fr
http://georessources.univ-lorraine.fr/



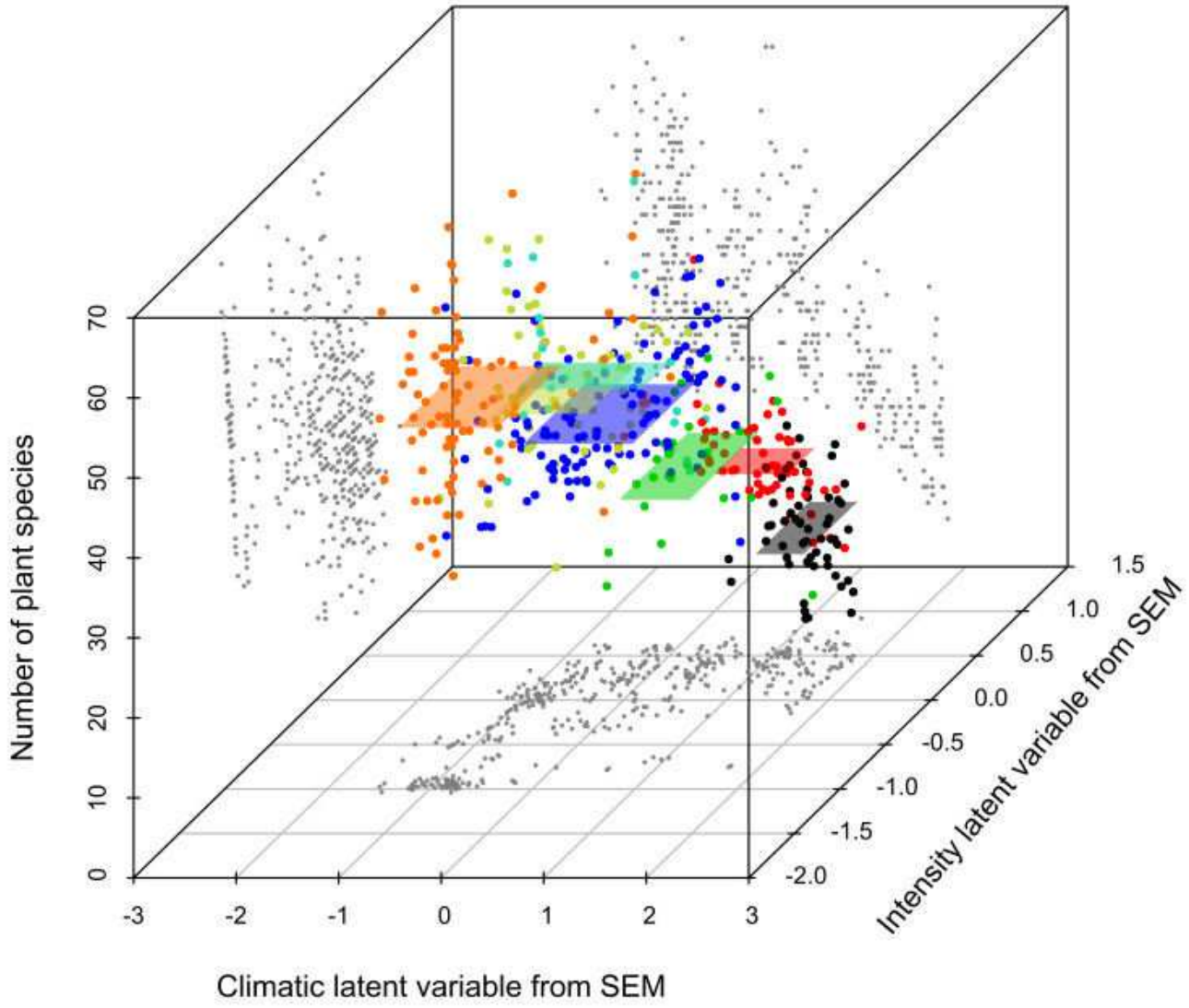
absence de liaison



forte liaison



3 groupes homogènes


Number of plant species
0
10
20
30
40
50
60
70
-3
-2
-1
0
1
2
3
Climatic latent variable from SEM
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
Intensity latent variable from SEM

statisitques multivariées = simplifier les données brutes pour pouvoir visualiser le lien entre les variables et repérer les macro-caractéristiques (analyses factorielles).

Plan du cours

1. Analyse en composantes principales
    1. Objectifs d'une ACP
    2. Principes de base
    3. Écriture matricielle
    4. Plans de projection
    5. Interprétation
        1. Matrice des corrélations
        2. Valeurs propres
        3. Vecteurs propres
        4. Représentations graphiques

Prérequis :

- Statistiques descriptives
- Calcul matriciel
- Vecteurs et projections

Plan du cours

1. Analyse en composantes principales
    1. Objectifs d'une ACP
    2. Principes de base
    3. Écriture matricielle
    4. Plans de projection
    5. Interprétation
        1. Matrice des corrélations
        2. Valeurs propres
        3. Vecteurs propres
        4. Représentations graphiques

Objectifs d'une ACP
1. Explorer et décrire un jeu de données multivarié
2. Visualiser les résultats dans un espace réduit
3. Identifier les groupes d'individus ou varibles ayant des comportements similaires
4. Déterminer les corrélations entre les variables

On peut travailler sur des données quantitatives ou catégorielles.
Marche bien si il y a un grand nombre d'individus.

La deuxième composante principale doit être orthogonale à la 1ère composante = Phénomènes indépendants et non corrélés selon la direction de plus grande inertie restante.
Le nombre de composante principal =
- soit k (nombre de variables)
- soit i - 1 (nombre d'individus -1)

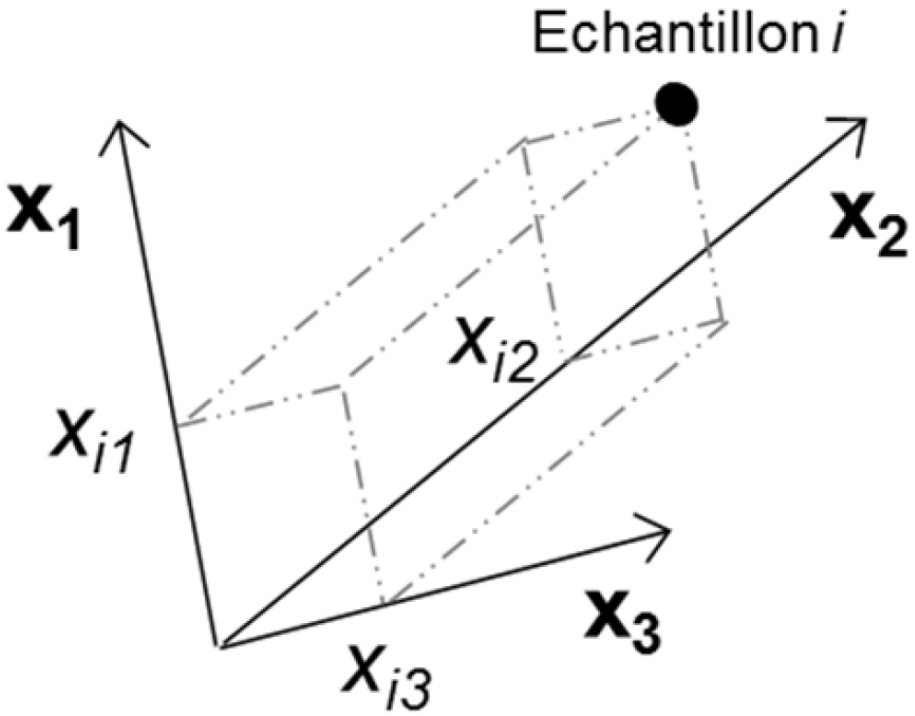
=> min (u i- 1)

2. principes de base

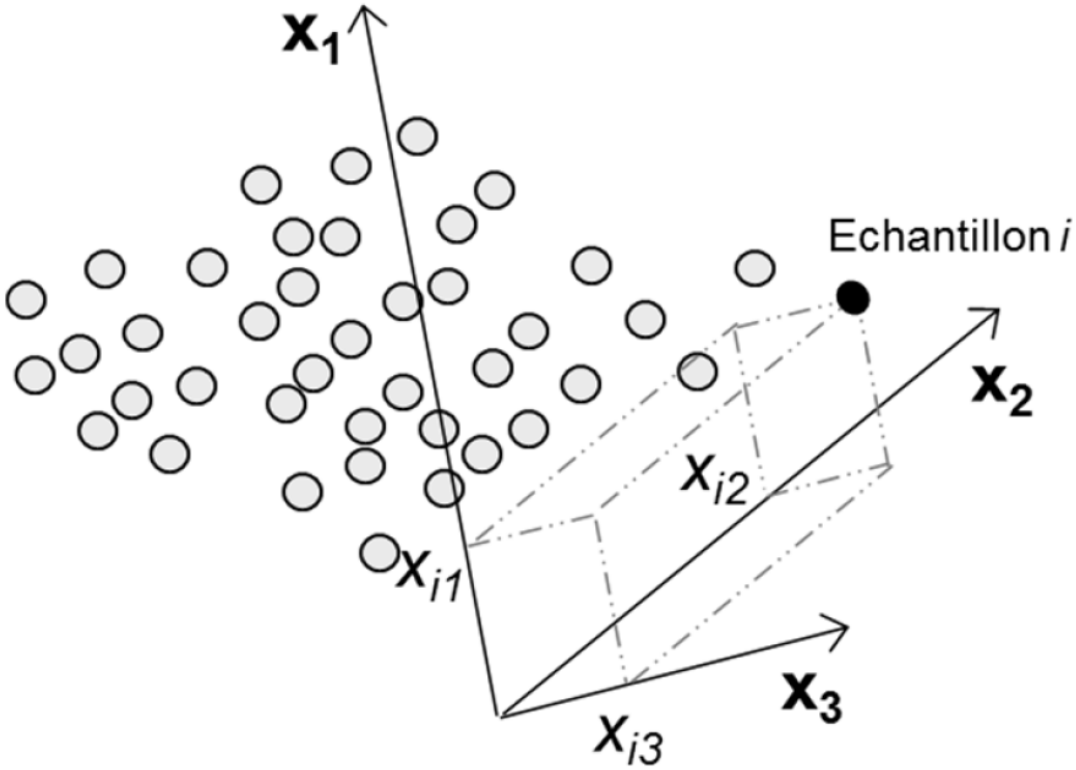
n individus
k variables

on va représenter sous forme de tableau avec n lignes et k colonnes.

Chaque individu est représenté dans l'espace des variables par un point à n coordonnées. Chaque variable est représentée dans lespace des individus par un point à n coordonnées.



Représentation de l'échantillon *i* dans l'espace des variables (x_1, x_2, x_3).

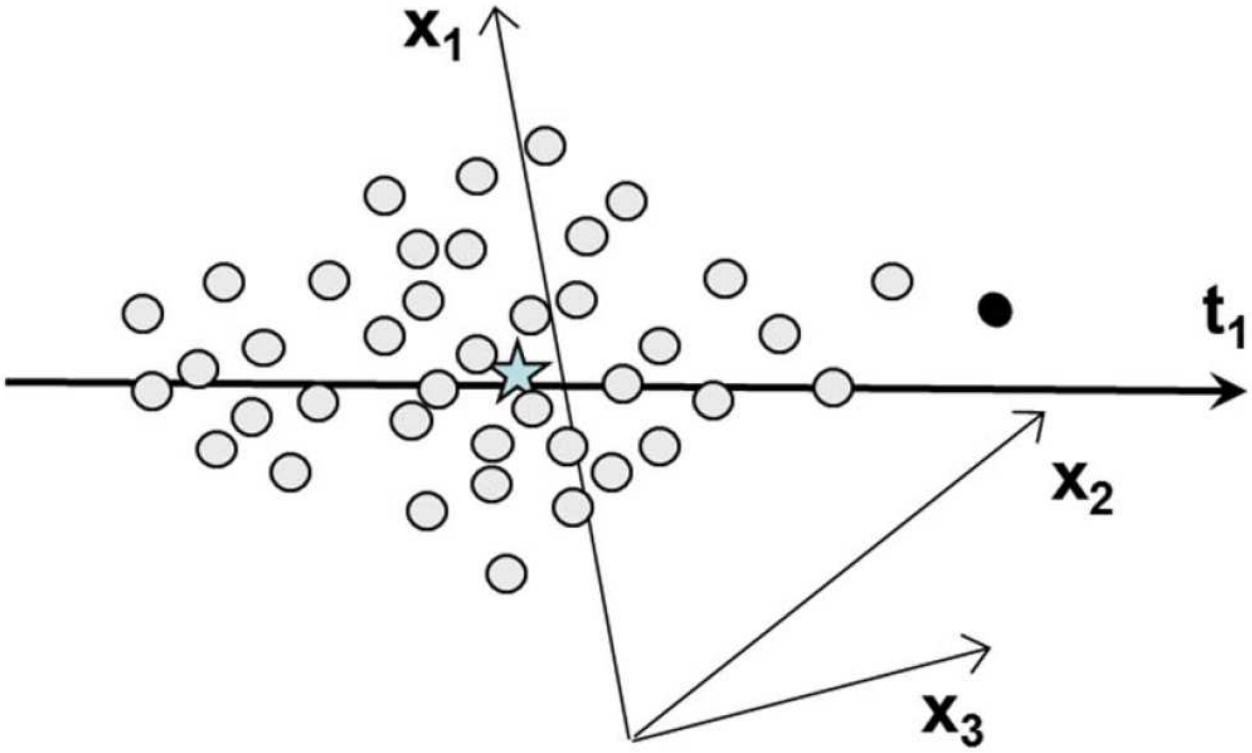


Représentation d'une population d'individus dans l'espace des variables (x_1, x_2, x_3).

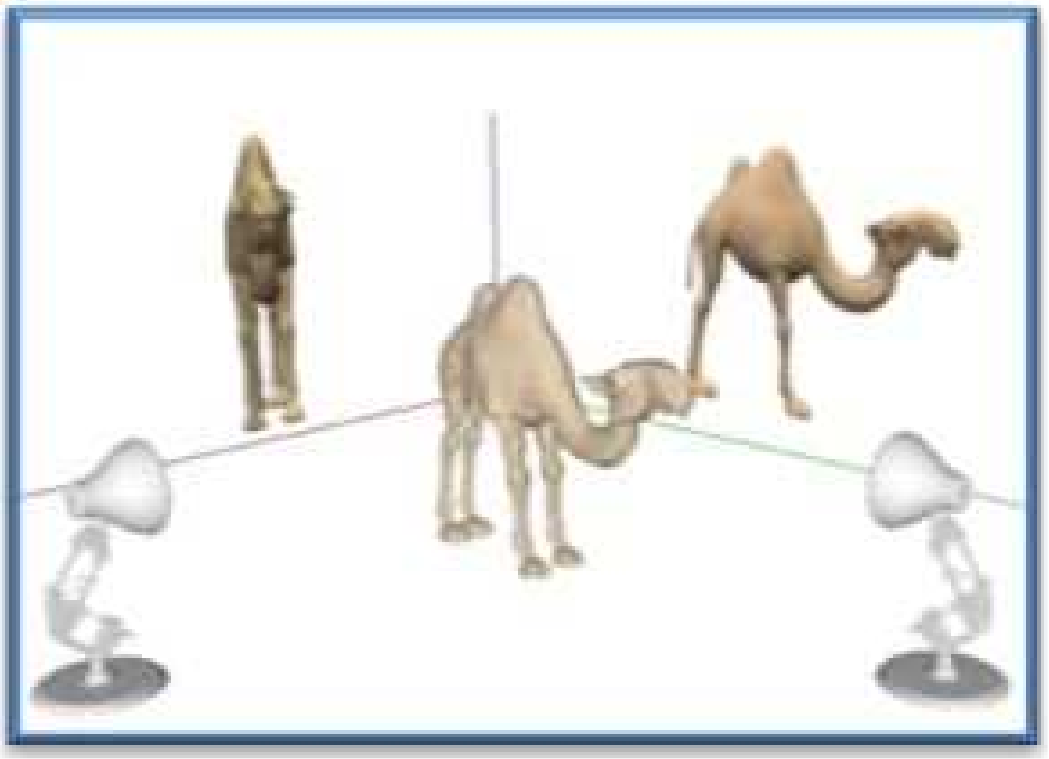
L'information totale contenue dans un nuage de point s'appelle l'inertie = somme de toutes les variables. On charche la direction selon laquelle l'inertie est maximale.

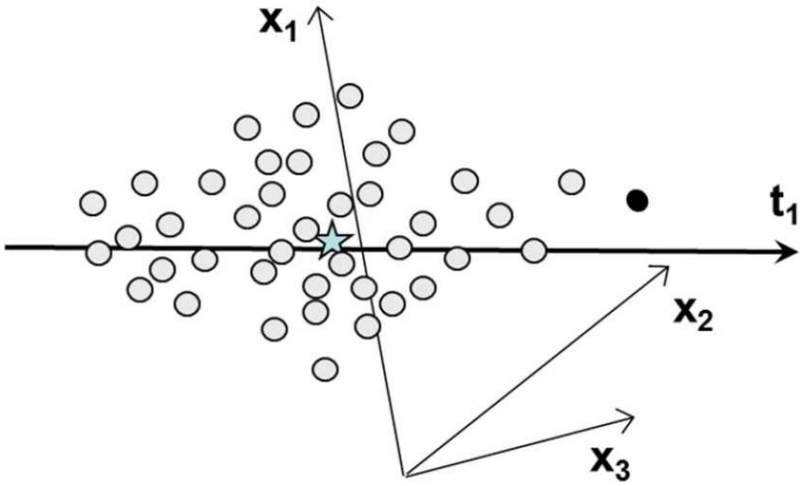
2. principes de base

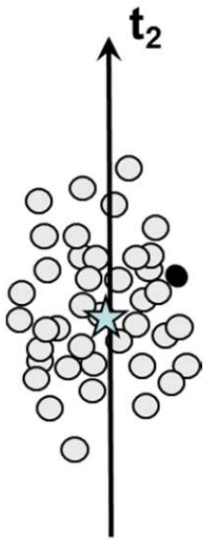
Chaque combinaison principale est une composante linéaire des variables initiales.



2. principes de base





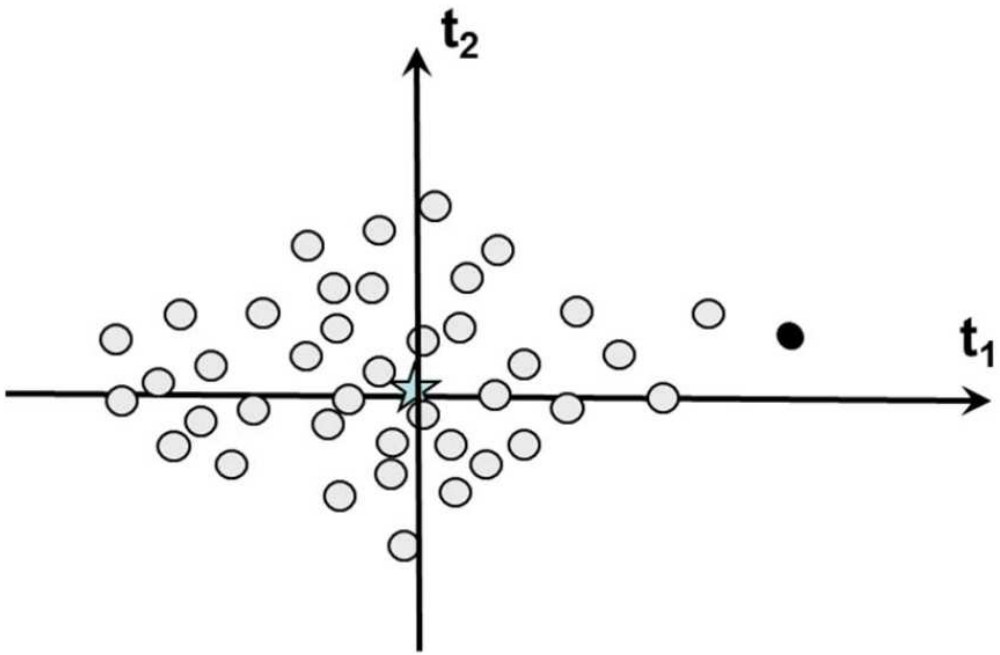


La deuxième composante principale doit être orthogonale à la 1ère composante = Phénomènes indépendants et non corrélés selon la direction de plus grande inertie restante.
Le nombre de composante principal =
- soit k (nombre de variables)
- soit i - 1 (nombre d'individus -1)

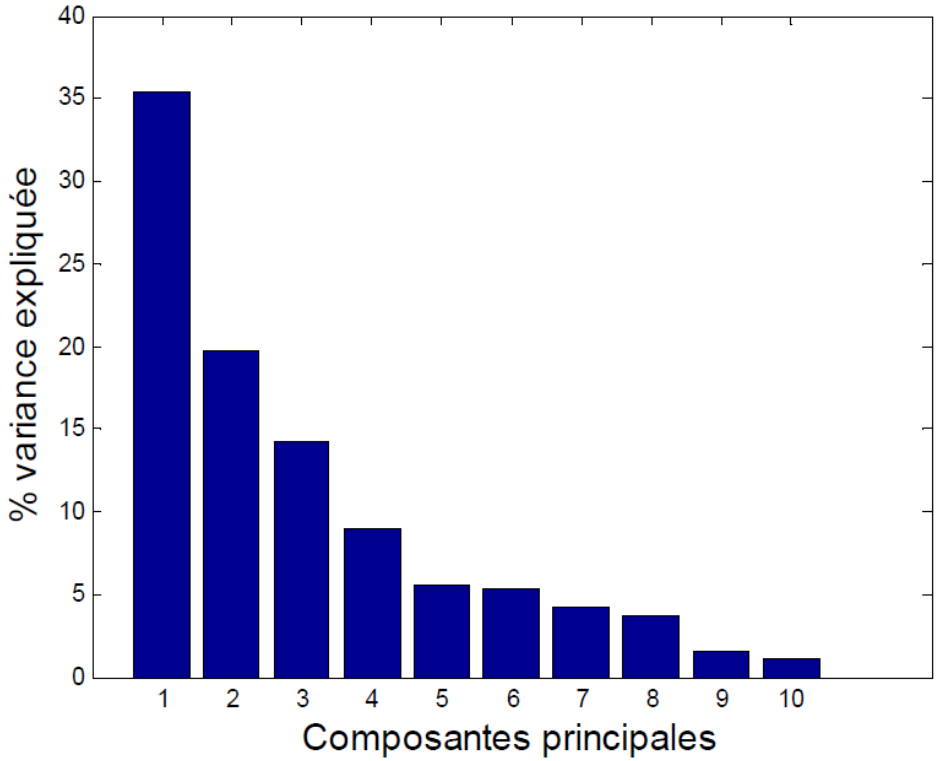
=> min k ou i- 1

2. principes de base

« score plot »
carte des individus projetés sur deux CP

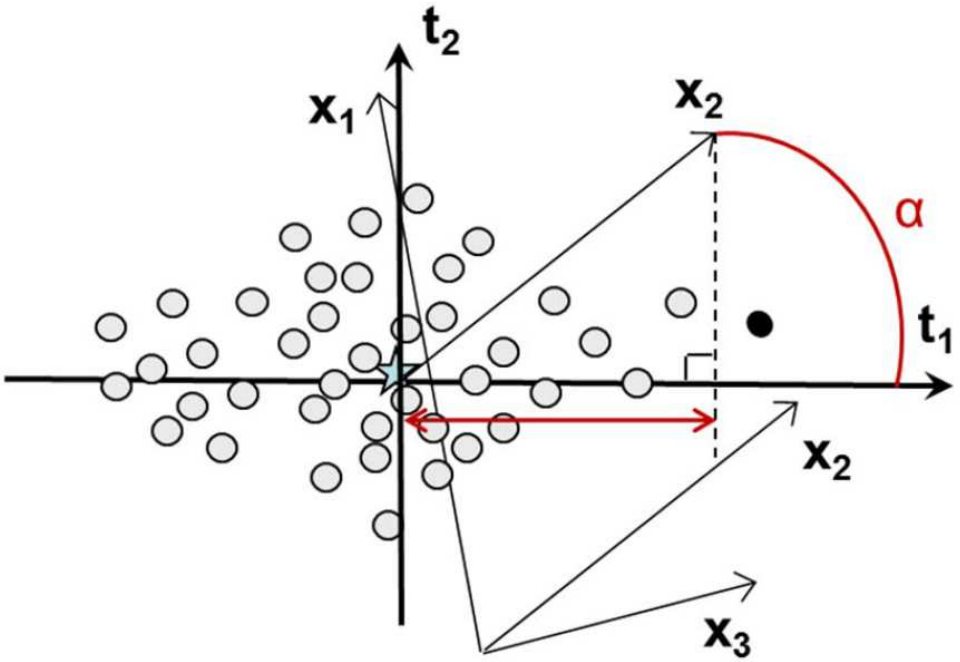


« scree plot »
diagramme des variances expliquées



2. principes de base

« loading plot »
carte des variables projetées sur deux CP



Plan du cours

1. Analyse en composantes principales
    1. Objectifs d’une ACP
    2. Principes de base
    3. Écriture matricielle
    4. Plans de projection
    5. Interprétation
        1. Matrice des corrélations
        2. Valeurs propres
        3. Vecteurs propres
        4. Représentations graphiques

Le principe de l'ACP est un changement de base. On part d'une matrice de données (n,k) vers une nouvelle matrice (kxA) telle que
- la variante des nouvelles variables et décroissante
- les nouvelles variables ne sont plus corrélées

Saisissez du texte ici

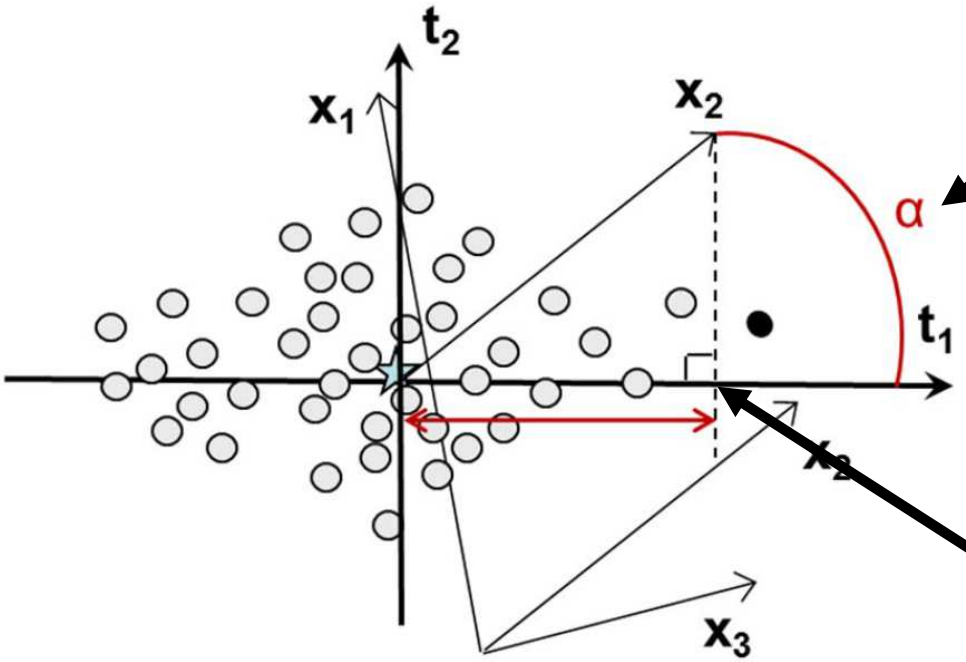
3. Ecriture matricielle

Chaque composante est une combinaison linéaire des variables initiales.

X = T . P' + t

avec X = matrice des données (n x k)
T= la matrice des scores (n x A)
P' = transposée de la matrice des loadings
t = transposée de la matrices des résidus

« loading plot »
carte des variables projetées sur deux CP



Corrélation entre les deux vecteurs

Contribution factorielle de la variable dans la CP

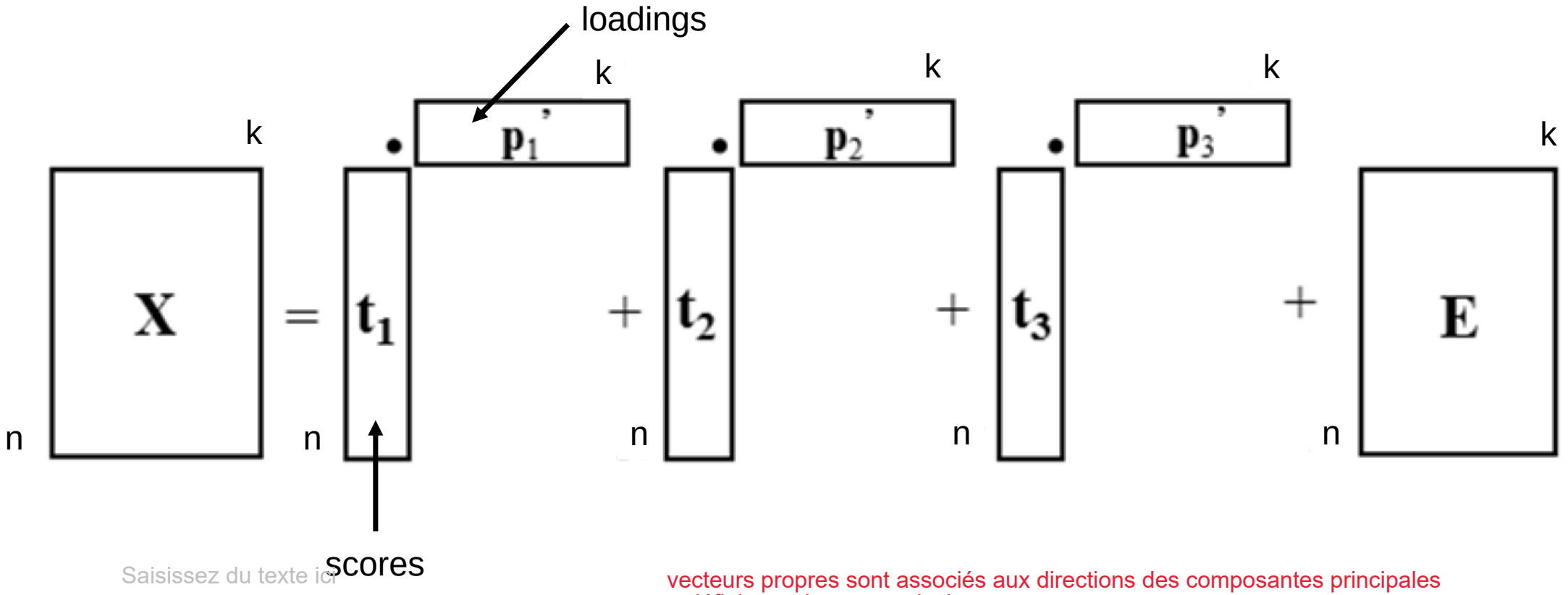
$$t_1 = p_{11}x_1 + p_{12}x_2 + \cdots + p_{1k}x_k$$

3. Ecriture matricielle

Comment calculer T et P ?
- Diagonaliser le produit scalaire X'.X
 = matrice des variances - covariances / des corrélations

$$X = T \bullet P' + E$$

Exemple pour 3 CP :



Saisissez du texte ici

vecteurs propres sont associés aux directions des composantes principales
 - définissent les axes principaux
 - sont orthogonaux entre eux

valeurs propres représentent l'inertie capturée par chaque composantes principales
 - variance de chaques composantes principales
 - toujours positives

3. Ecriture matricielle

	variable x1	variable x2
individu 1	3	2
individu 2	4	10
individu 3	6	5
individu 4	6	8
individu 5	6	10
individu 6	7	2
individu 7	7	13
individu 8	8	9
individu 9	9	5
individu 10	9	8
individu 11	9	14
individu 12	10	7
individu 13	11	12
individu 14	12	10
individu 15	12	11
individu 16	13	6
individu 17	13	14
individu 18	13	15
individu 19	13	17
individu 20	14	7
individu 21	15	13
individu 22	17	13
individu 23	17	17
individu 24	18	19
individu 25	20	20

Exemple avec deux variables x_1 et x_2, 25 individus

var(x1) = 20,277 var(x2) = 24,060 Cov(x1,x2) = 15,585

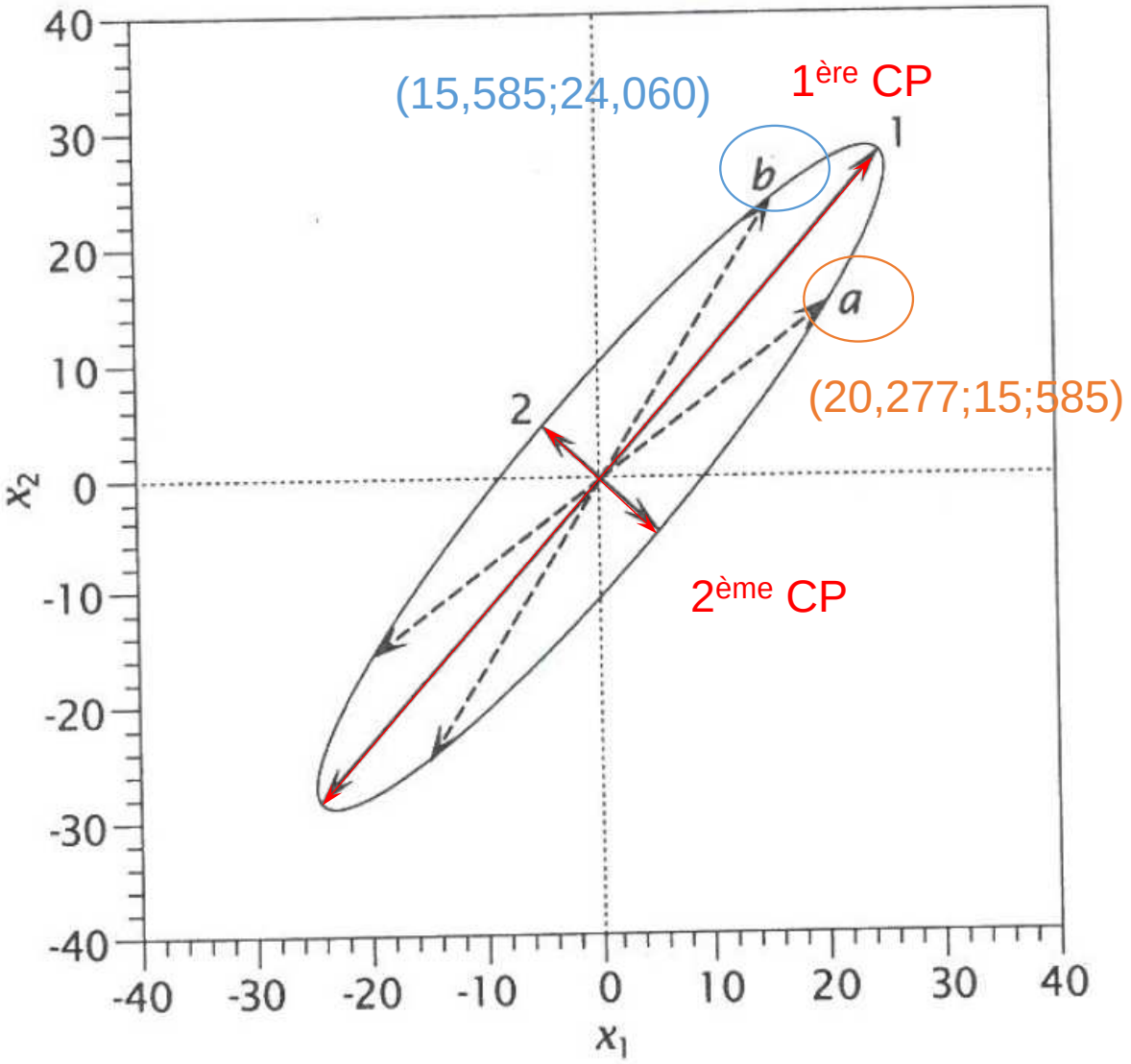
$$S = \begin{bmatrix} 20{,}277 & 15{,}585 \\ 15{,}585 & 24{,}060 \end{bmatrix}$$

Matrice de variances/covariances

Permet de définir deux vecteurs sur l'espace des variables :

(20,277;15;585) (15,585;24,060)

3. Ecriture matricielle



$$S = \begin{bmatrix} 20{,}277 & 15{,}585 \\ 15{,}585 & 24{,}060 \end{bmatrix}$$

Valeurs propres
Vecteurs propres

$$U_1 = \begin{bmatrix} 0.663 \\ 0.748 \end{bmatrix} \qquad U_2 = \begin{bmatrix} -0.748 \\ 0.663 \end{bmatrix}$$

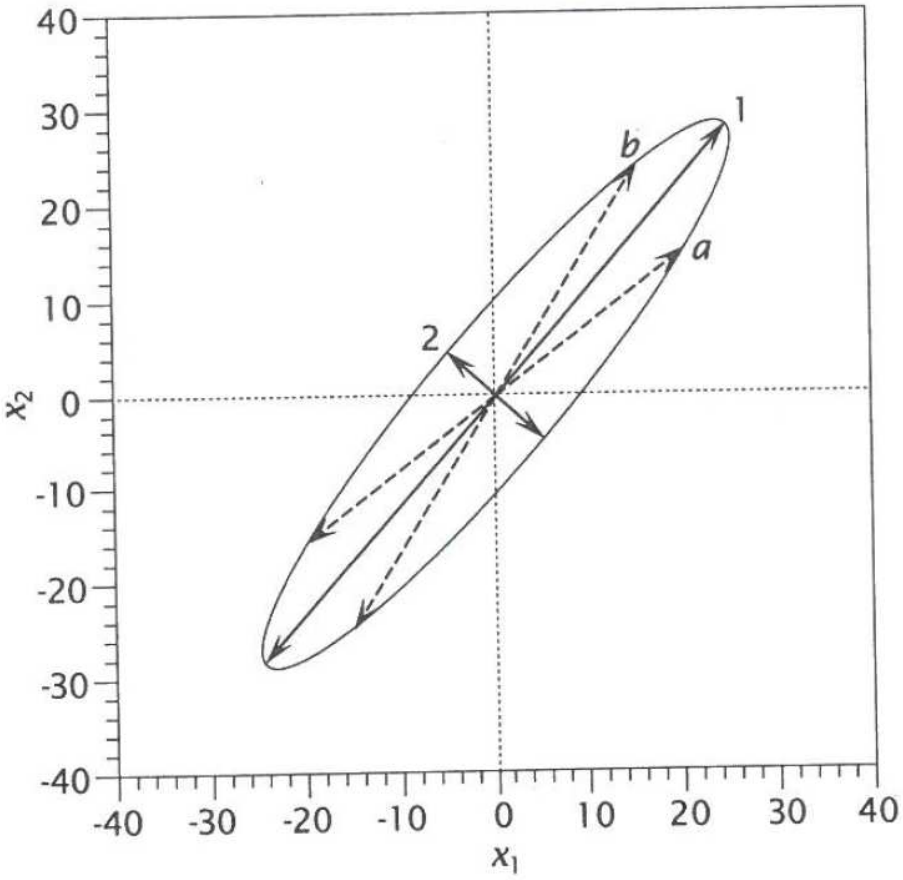
1er vecteur propre — 2ème vecteur propre

37.868 — 6.469

1ère valeur propre — 2ème valeur propre

Longueur des axes = valeurs propres

3. Ecriture matricielle



var(x1) = 20,277 ; var(x2) = 24,060

$$S = \begin{bmatrix} 20{,}277 & 15{,}585 \\ 15{,}585 & 24{,}060 \end{bmatrix}$$

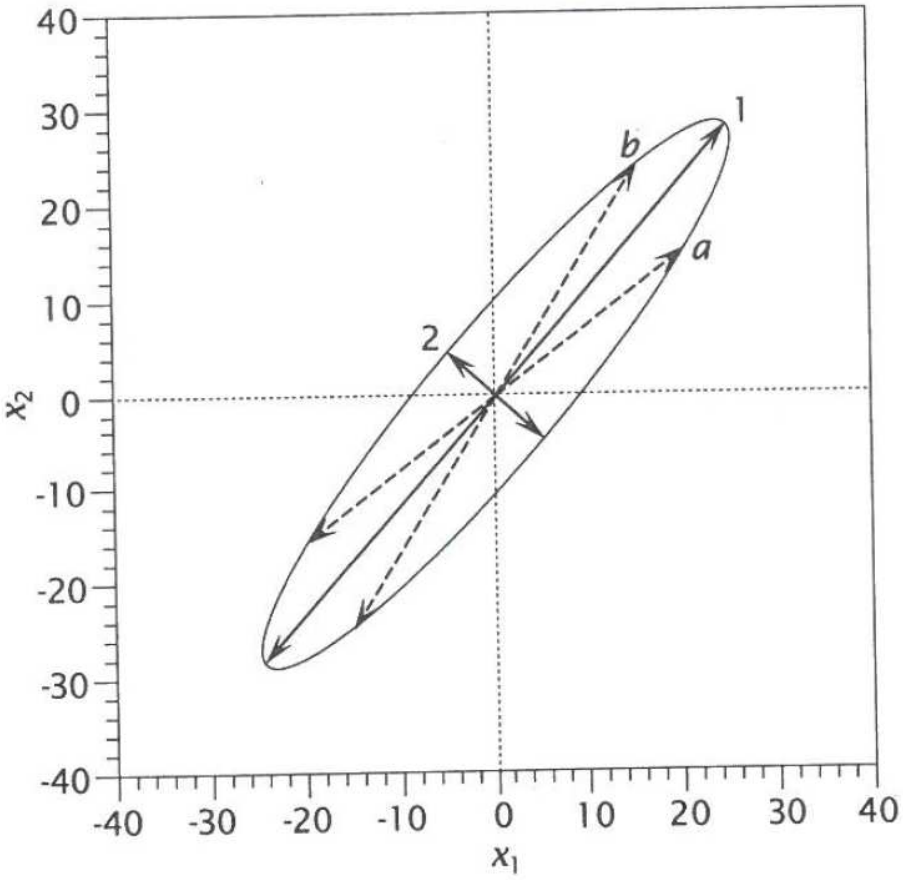
$$U_1 = \begin{bmatrix} 0.663 \\ 0.748 \end{bmatrix} \qquad U_2 = \begin{bmatrix} -0.748 \\ 0.663 \end{bmatrix}$$

37.868 6.469

Variance totale = trace S
= somme des valeurs propres
= var(x1)+var(x2)
= 44.337

La première CP contient
37.868/44.337 = 86% de la variance totale.
La deuxième CP contient 14% de la variance totale.

3. Ecriture matricielle



$$S = \begin{bmatrix} 20{,}277 & 15{,}585 \\ 15{,}585 & 24{,}060 \end{bmatrix}$$

$$U_1 = \begin{bmatrix} 0.663 \\ 0.748 \end{bmatrix} \qquad U_2 = \begin{bmatrix} -0.748 \\ 0.663 \end{bmatrix}$$

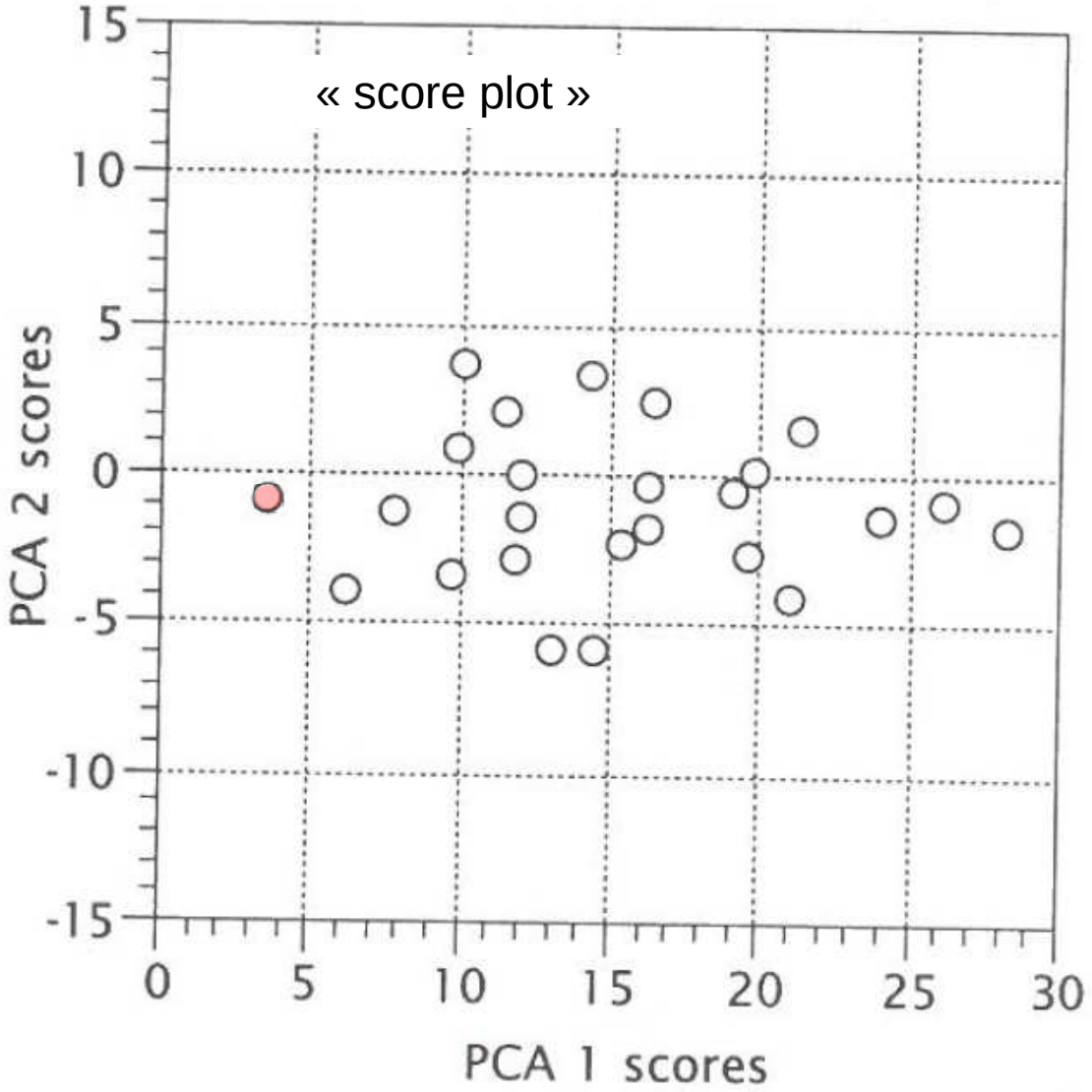
En calculant S^R= XU, on obtient les « scores » des observations sur chacune des CP.

$$s_{i1} = 0.663x_{i1} + 0.748x_{i2}$$
$$s_{i2} = -0.748x_{i1} + 0.663x_{i2}$$

= projection des individus sur les deux axes de l'ACP.
= scores des individus/observations

3. Ecriture matricielle

PCA 1	PCA 2	PCA 1	PCA 2
3.485	−0.918	15.44	−2.346
10.130	3.638	16.18	−1.683
7.718	−1.173	13.11	−5.746
9.962	0.816	19.09	−0.442
11.460	2.142	19.84	0.221
6.137	−3.910	21.34	1.547
14.370	3.383	14.52	−5.831
12.040	−0.017	19.67	−2.601
9.707	−3.417	21.00	−4.097
11.950	−1.428	23.99	−1.445
16.440	2.550	26.15	−0.867
11.870	−2.839	28.22	−1.700
16.270	−0.272		



Plan de projection : représentation des individus ou des variables dans un espace à deux dimensions -> on effectue une projection des n dimensions sur un plan (2D).

Plan de projection idéal = plan factoriel (minimise les distorsions)

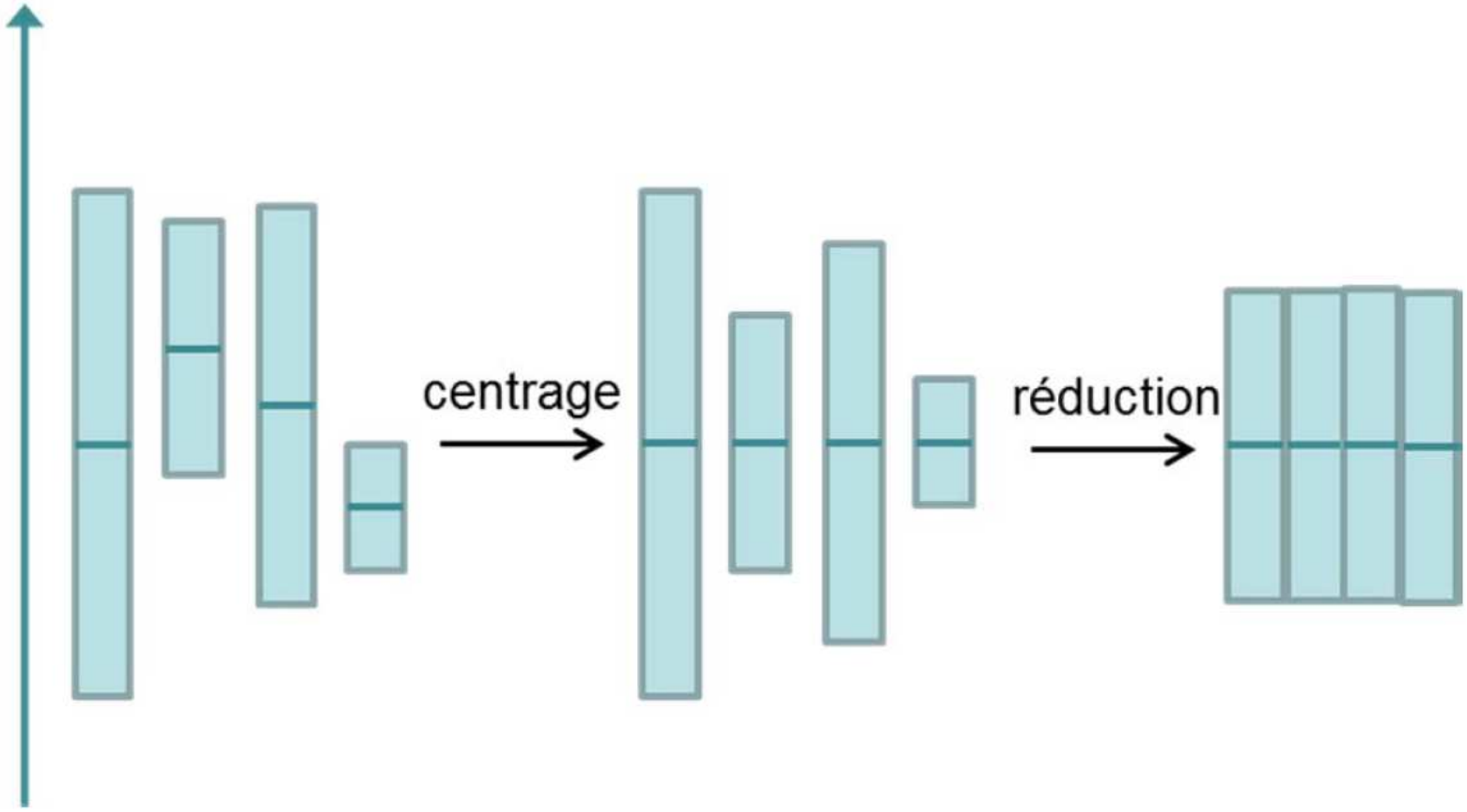
Plan central = plan factoriel qui contient le maximum de variablitié

Plan du cours

1. Analyse en composantes principales
    1. Objectifs d'une ACP
    2. Principes de base
    3. Écriture matricielle
    4. Plans de projection
    5. Interprétation
        1. Matrice des corrélations
        2. Valeurs propres
        3. Vecteurs propres
        4. Représentations graphiques

Plus de poids pour les variables dont la variance plus élevée
- + bruits
- plus l'échelle est + grande

5. Interprétation



Exemple

Compilation des paramètres d'une série d'échantillons de pétrole brut dont la source est connue.

No.	API Gravity	Sulfur, %	Pr/Ph	SAT/ARO	Oil CIR	Gasoline CIR	C G-R	rock	
1	24.6	1.69	1.1	1.1	-26.23	-26.3	-0.27	carbonate	Carbonate, Deltaic, Marine Shale
2	27	1.58	0.95	1.1	-26.62	-26.89	-0.33	carbonate	
3	28.1	1.53	1.02	1.2	-26.02	-26.21	-0.39	carbonate	API gravity
4	29.5	3.1	0.7	0.8	-26.1	-27.16	-1.42	carbonate	Pr/Ph = pristane/phytane ratio;
5	32.2	2.61	0.65	0.8	-26.24	-27.2	-1.09	carbonate	SAT/ARO = saturates to aromatics ratio;
6	33.6	2.27	0.75	0.7	-26.5	-27.19	-0.93	carbonate	Oil CIR = whole-oil carbon isotope ratio;
7	31.7	2.52	0.7	0.9	-26.24	-27.07	-1.12	carbonate	Gasoline CIR = carbon isotope ratio of gasoline fraction;
8	33	1.71	0.71	1.2	-26.27	-27	-0.97	carbonate	C G-R = difference in carbon isotope ratio between gasoline fraction and residuum.
9	34	1.95	0.62	1.2	-26.3	-26.95	-0.96	carbonate	From Chung, et al, 1994, Table 1.
10	28	2.78	0.67	0.7	-26.57	-27.46	-0.83	carbonate	
11	25.5	2.26	0.82	0.9	-25.59	-25.8	-0.6	carbonate	
12	35.4	1.03	0.85	1.3	-25.25	-25.65	-0.5	carbonate	
13	35.1	1.39	0.58	1.1	-25.06	-25.52	-0.54	carbonate	

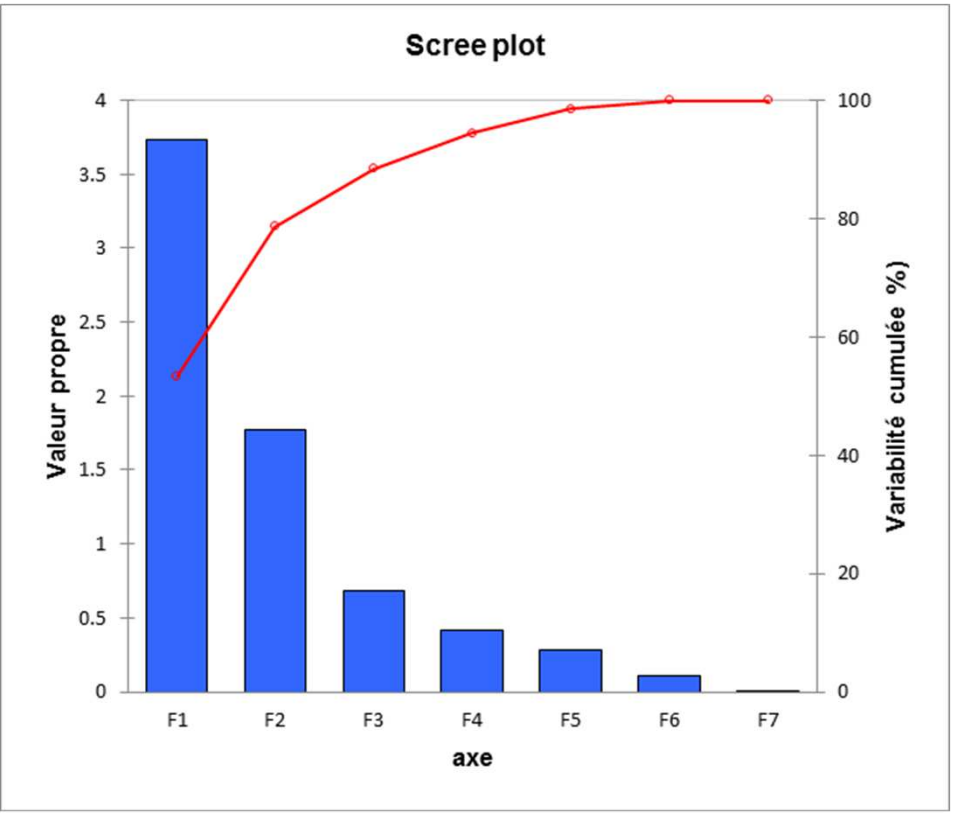
63 échantillons, 3 sources, 6 variables

5.1 matrice de corrélation

Corrélation ente les variables deux à deux

Variables	API Gravity	Sulfur, %	Pr/Ph	SAT/ARO	Oil CIR	Gasoline CIR	C G-R
API Gravity	**1**	-0.674	0.615	0.663	0.298	0.472	0.581
Sulfur, %	-0.674	**1**	-0.622	-0.466	0.124	-0.195	-0.693
Pr/Ph	0.615	-0.622	**1**	0.334	0.084	0.353	0.641
SAT/ARO	0.663	-0.466	0.334	**1**	0.202	0.467	0.713
Oil CIR	0.298	0.124	0.084	0.202	**1**	0.890	-0.103
Gasoline CIR	0.472	-0.195	0.353	0.467	0.890	**1**	0.342
C G-R	0.581	-0.693	0.641	0.713	-0.103	0.342	**1**

5.2 valeurs propres



Histogramme des valeurs propres
« scree plot »

F1 : 53% de variabilité expliquée par F1

F1+F2 = 79% : 79% de la variabilité totale des données peut être expliquée par une représentation binaire F1-F2

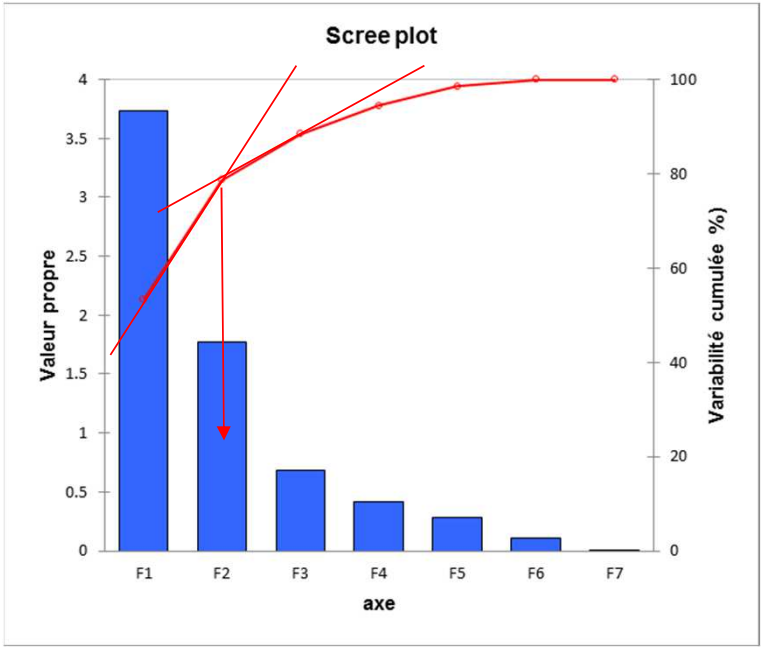
objectif = avoir une information concentrée sur un minimum d'axes

5.2 valeurs propres

règle de sélection pour déterminer les facteurs

- choix d'une valeur d'arrêt du pourcentage de variance cumulée (80%)
- scree test (Cattell, 1966) étude de la courbe croissante des valeurs propres

→ nombre de facteurs à retenir - premier point d'inflexion détecté sur la courbe.



- Règle de Kaiser (le facteur explique plus qu'une variable): valeur propre > 1

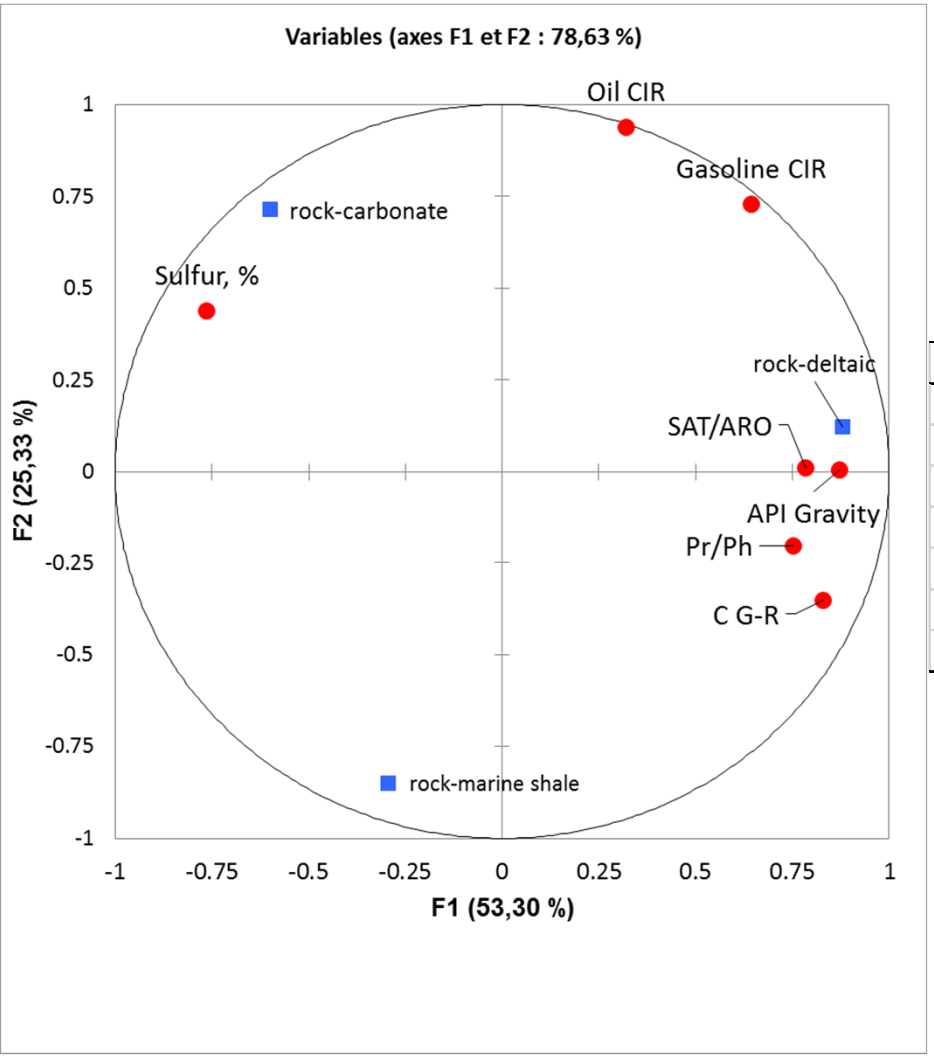
5.3 vecteurs propres

Contribution de chaque variable dans chaque composante/facteur

	F1	F2	F3	F4	F5	F6	F7
API Gravity	0.451	0.003	-0.042	-0.642	0.395	0.462	-0.117
Sulfur, %	-0.395	0.330	0.197	0.357	0.694	0.287	-0.074
Pr/Ph	0.390	-0.153	-0.633	0.341	0.439	-0.338	0.002
SAT/ARO	0.406	0.007	0.703	0.008	0.242	-0.531	-0.011
Oil CIR	0.166	0.705	-0.120	-0.082	-0.065	-0.067	0.667
Gasoline CIR	0.334	0.548	-0.045	0.245	-0.290	0.088	-0.659
C G-R	0.429	-0.266	0.218	0.526	-0.153	0.544	0.318

Force de la contribution des variables initiales dans les axes principaux

5.4 représentations graphiques des variables



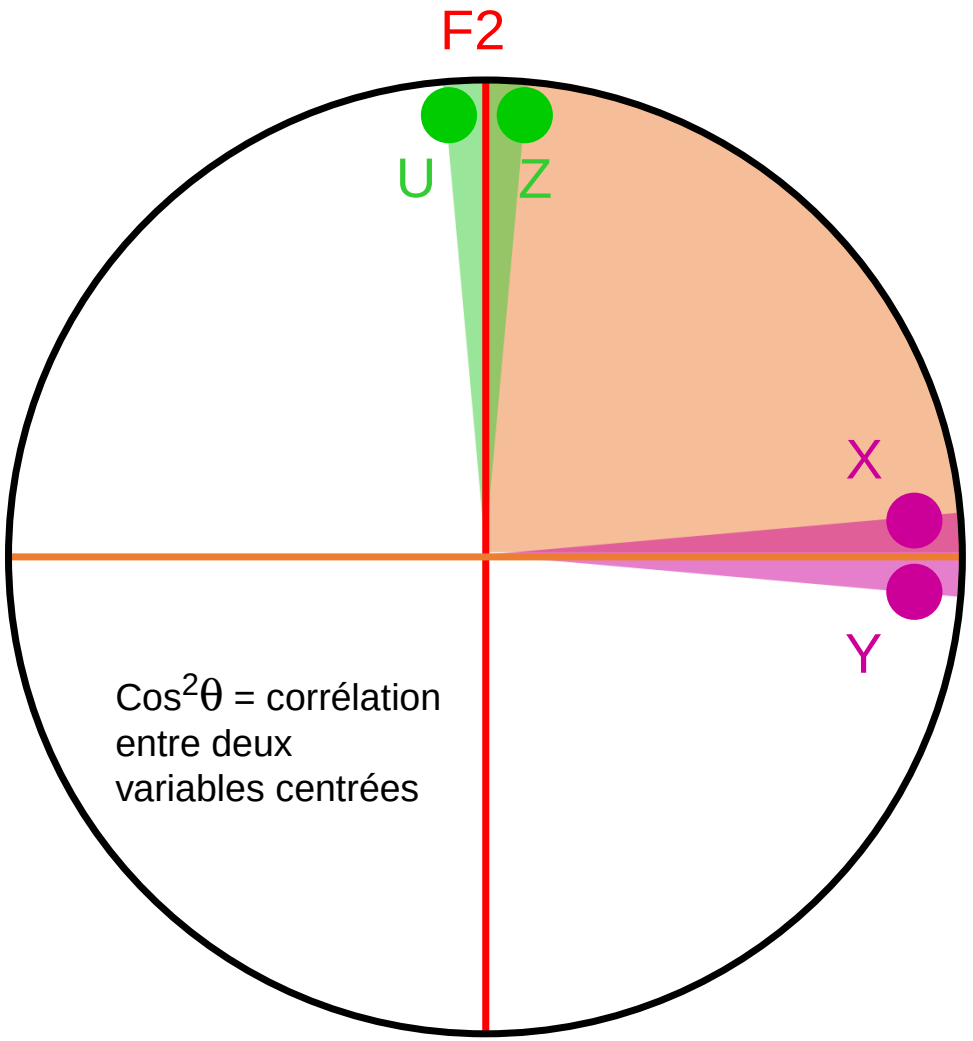
Cercle de corrélation sur les variables
« loading plot »

Coordonnées des variables :

	F1	F2	F3	F4	F5	F6	F7
API Gravity	0.871	0.004	-0.035	-0.417	0.210	0.149	-0.006
Sulfur, %	-0.763	0.439	0.163	0.232	0.369	0.092	-0.004
Pr/Ph	0.754	-0.204	-0.524	0.222	0.234	-0.109	0.000
SAT/ARO	0.784	0.010	0.582	0.005	0.129	-0.171	-0.001
Oil CIR	0.321	0.939	-0.099	-0.053	-0.035	-0.022	0.034
Gasoline CIR	0.645	0.729	-0.038	0.159	-0.155	0.028	-0.034
C G-R	0.830	-0.354	0.180	0.342	-0.082	0.175	0.016

5.4 représentations graphiques des variables

Cercle de corrélation sur les variables

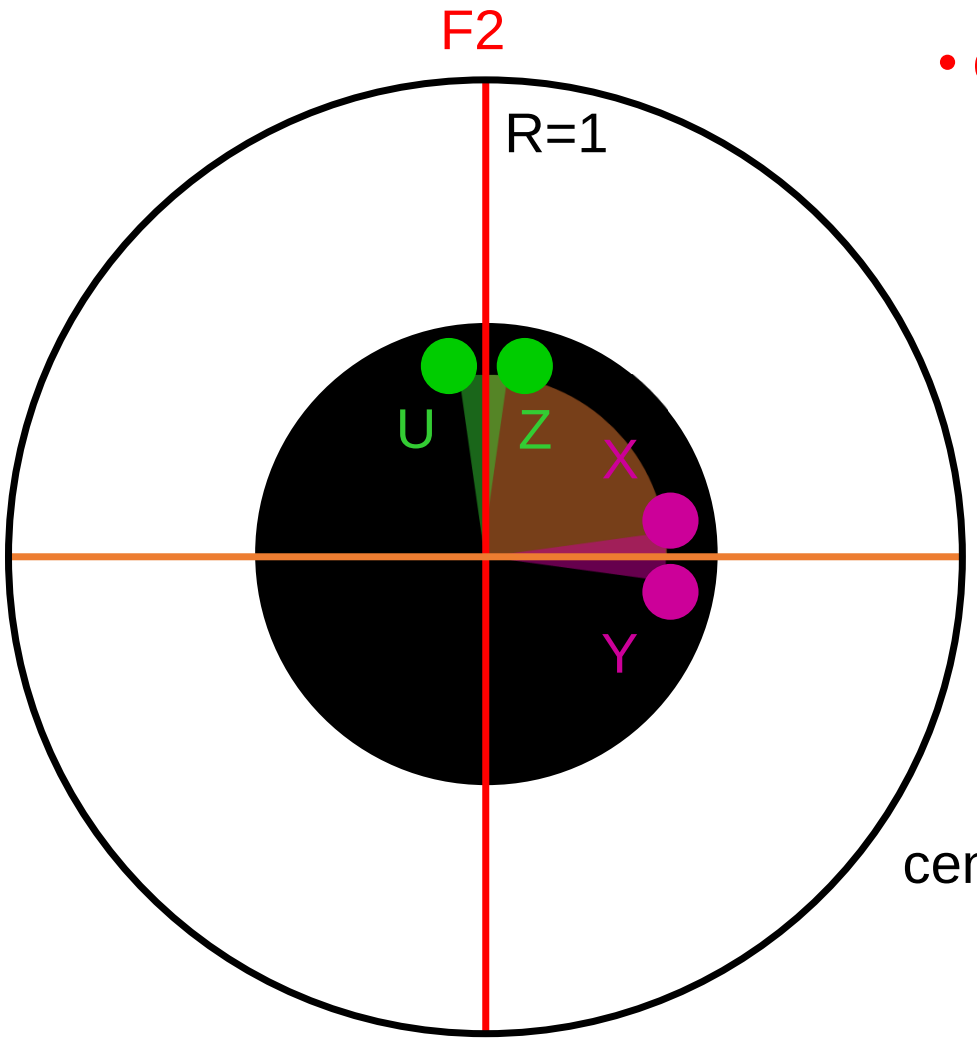


• corrélations positives R ≅ 1

1. X et Y sont fortement corrélées
2. U et Z sont fortement corrélées
3. X, Y et U,V sont totalement indépendantes
4. les qualités de corrélations sont valables

qualité de la représentation de chaque variable :
Cos^2 de l'angle entre la projection du vecteur et le facteur

5.4 représentations graphiques des variables

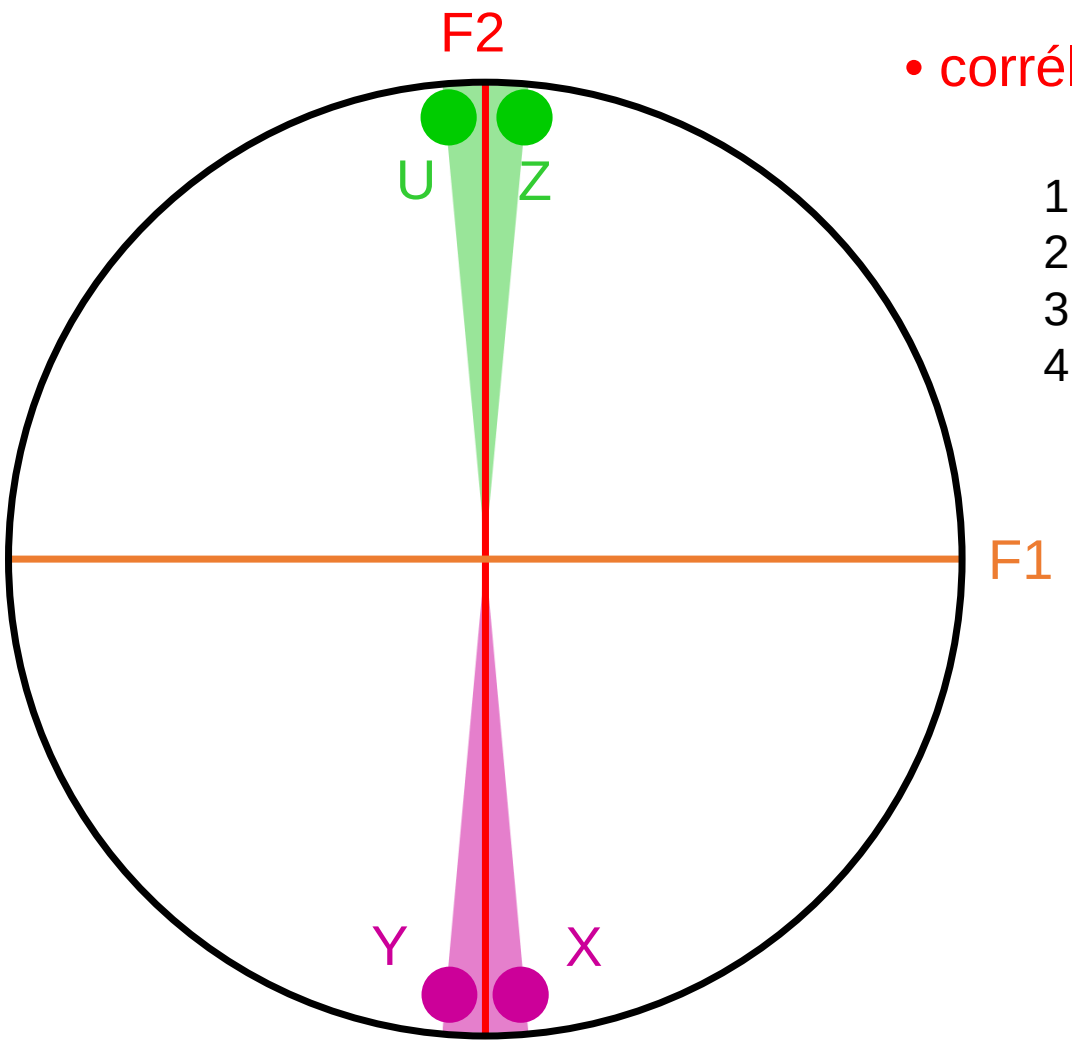


- corrélations positives R < 1

conclusions sur les corrélations risquent d'être erronées

centre du cercle des corrélations non interprétable

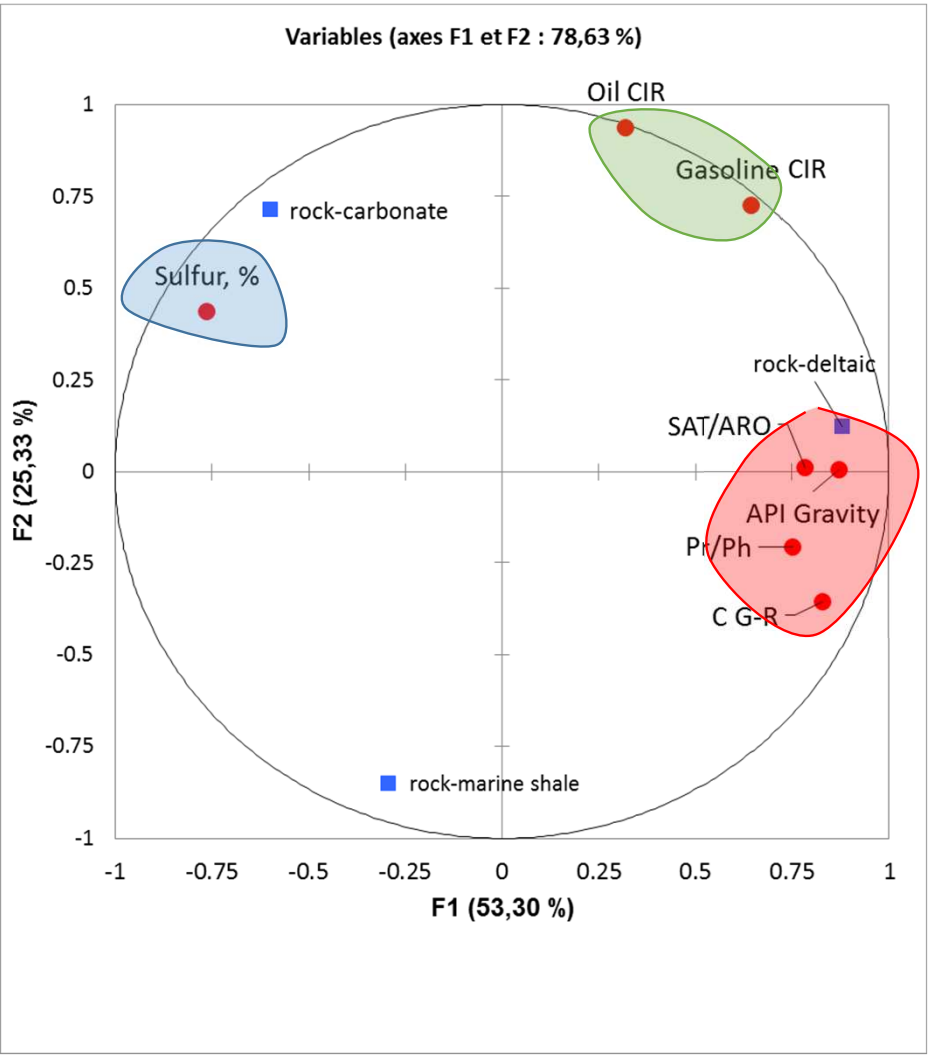
5.4 représentations graphiques des variables



- corrélations négatives R ≅ -1

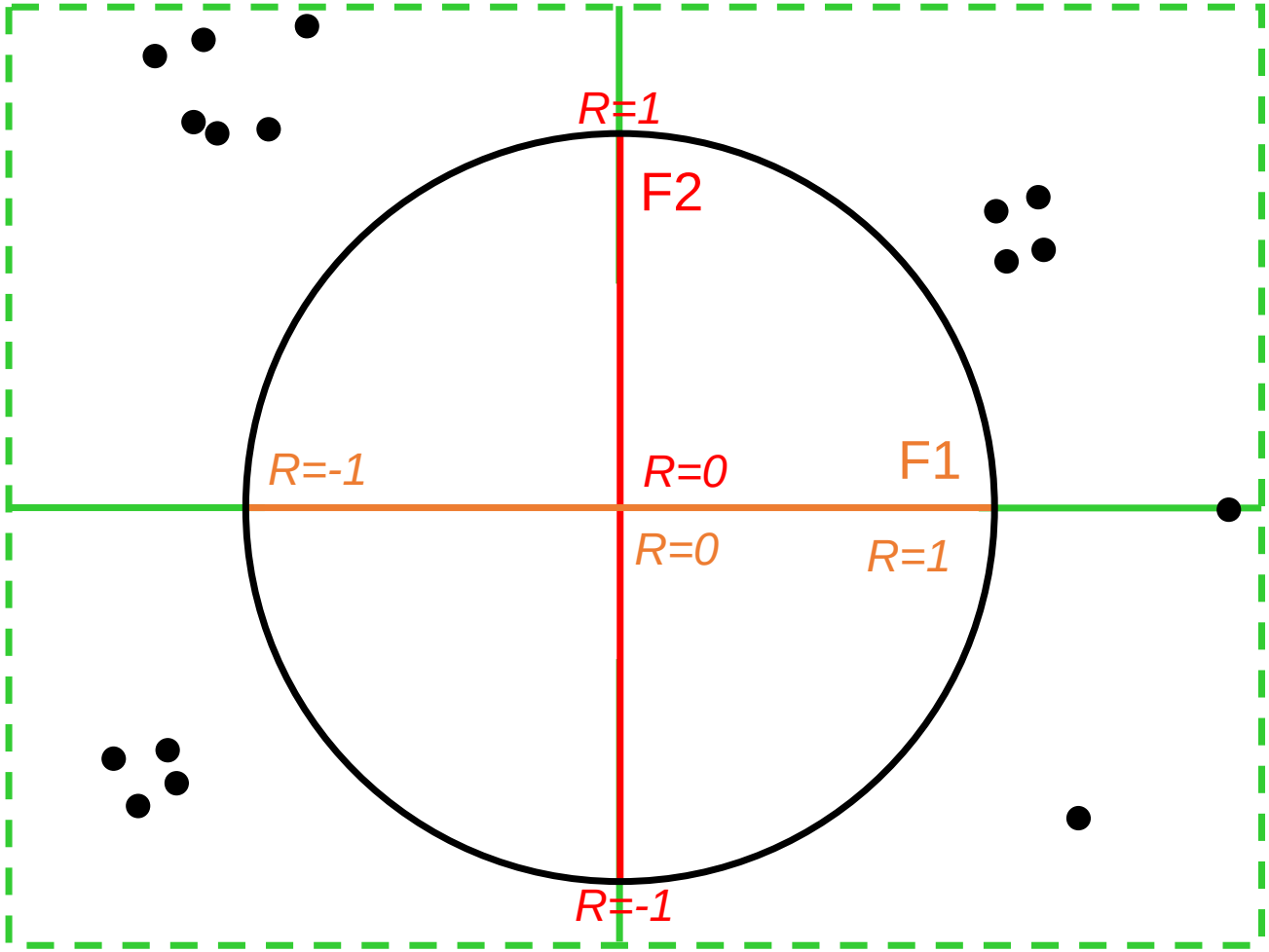
1. X et Y sont fortement corrélées
2. U et Z sont fortement corrélées
3. X, Y et U,V sont anti-corrélées
4. les qualités de corrélations sont valables

5.4 représentations graphiques des variables



Représentation graphique des variables corrélées
Deltaic ↔ groupe rouge
Carbonate ↔ ± sulfur%
Marine shale ↔ anti-corrélation groupe vert

5.4 représentations graphiques des observations

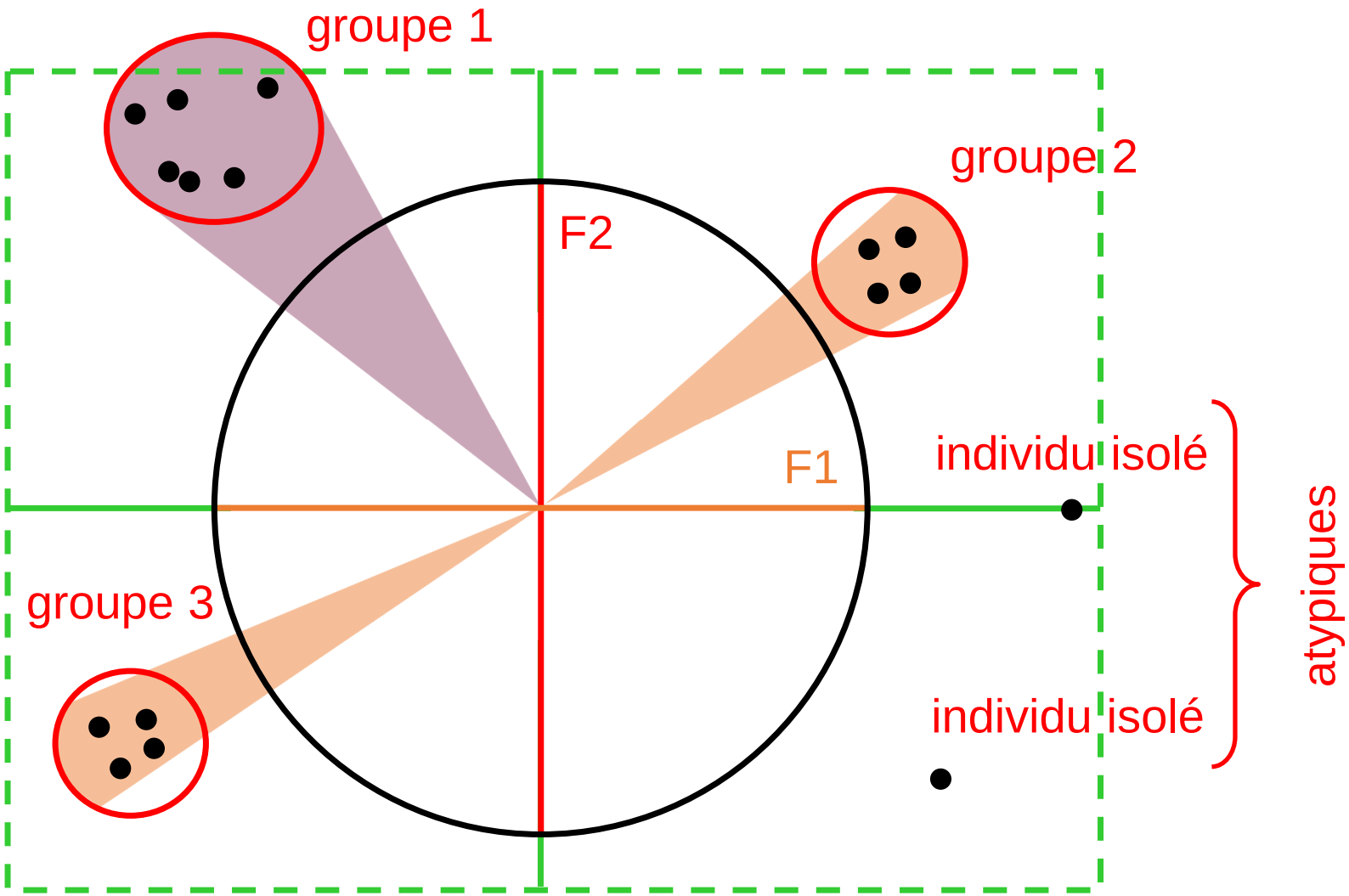


unités du cercle de corrélations
mais coordonnées différentes !!!

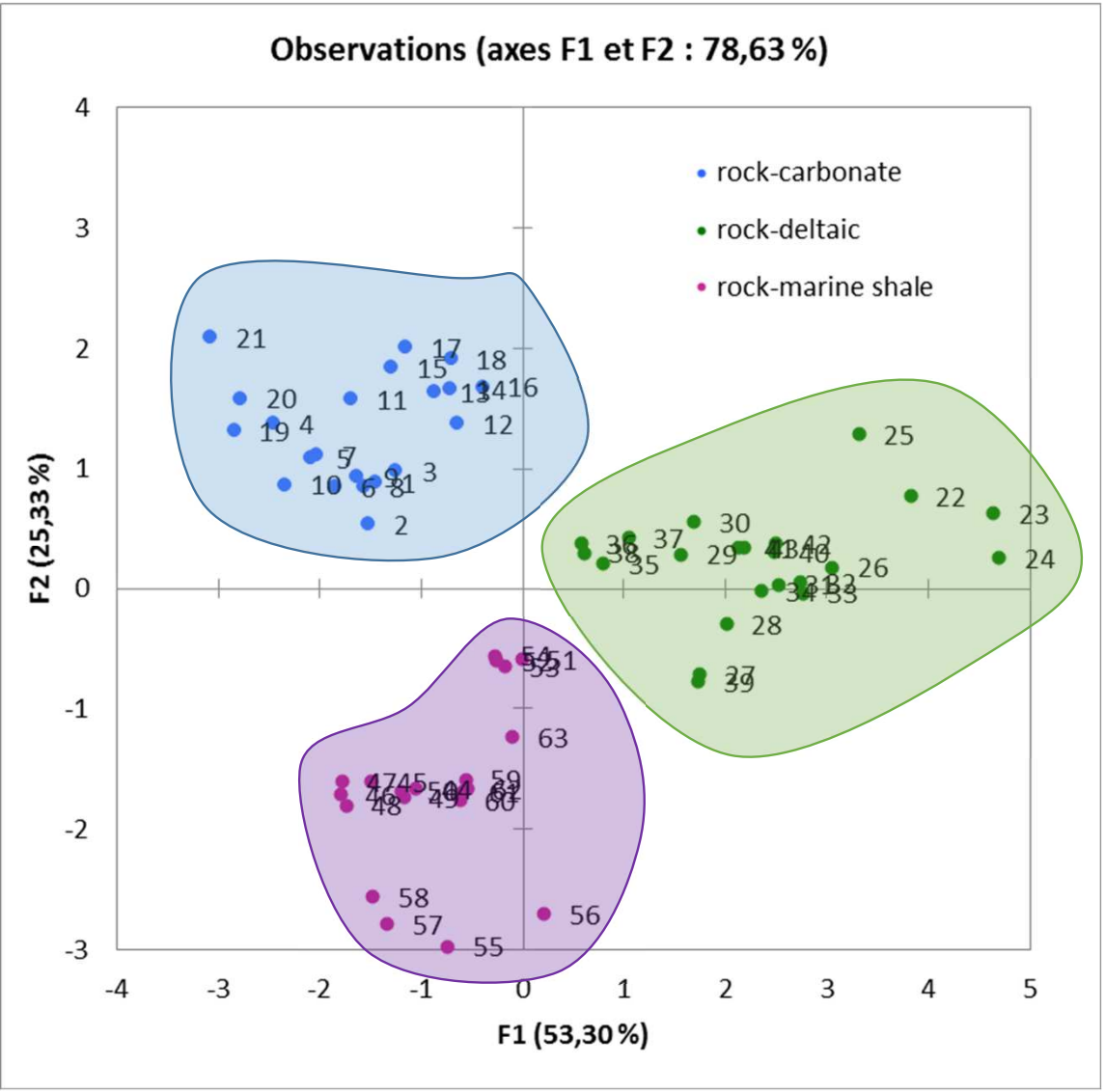
interprétations identiques

« score plot »

5.4 représentations graphiques des observations

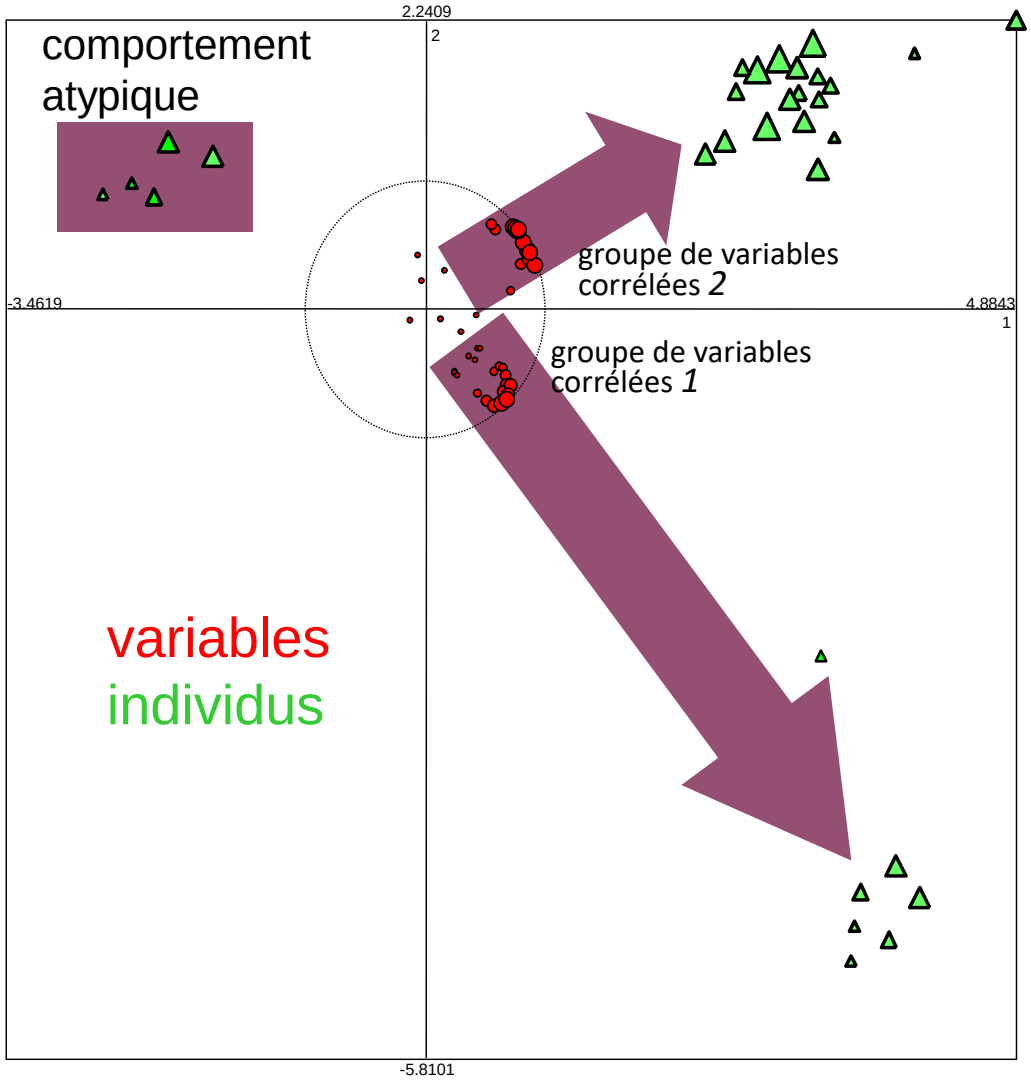


5.4 représentations graphiques des observations

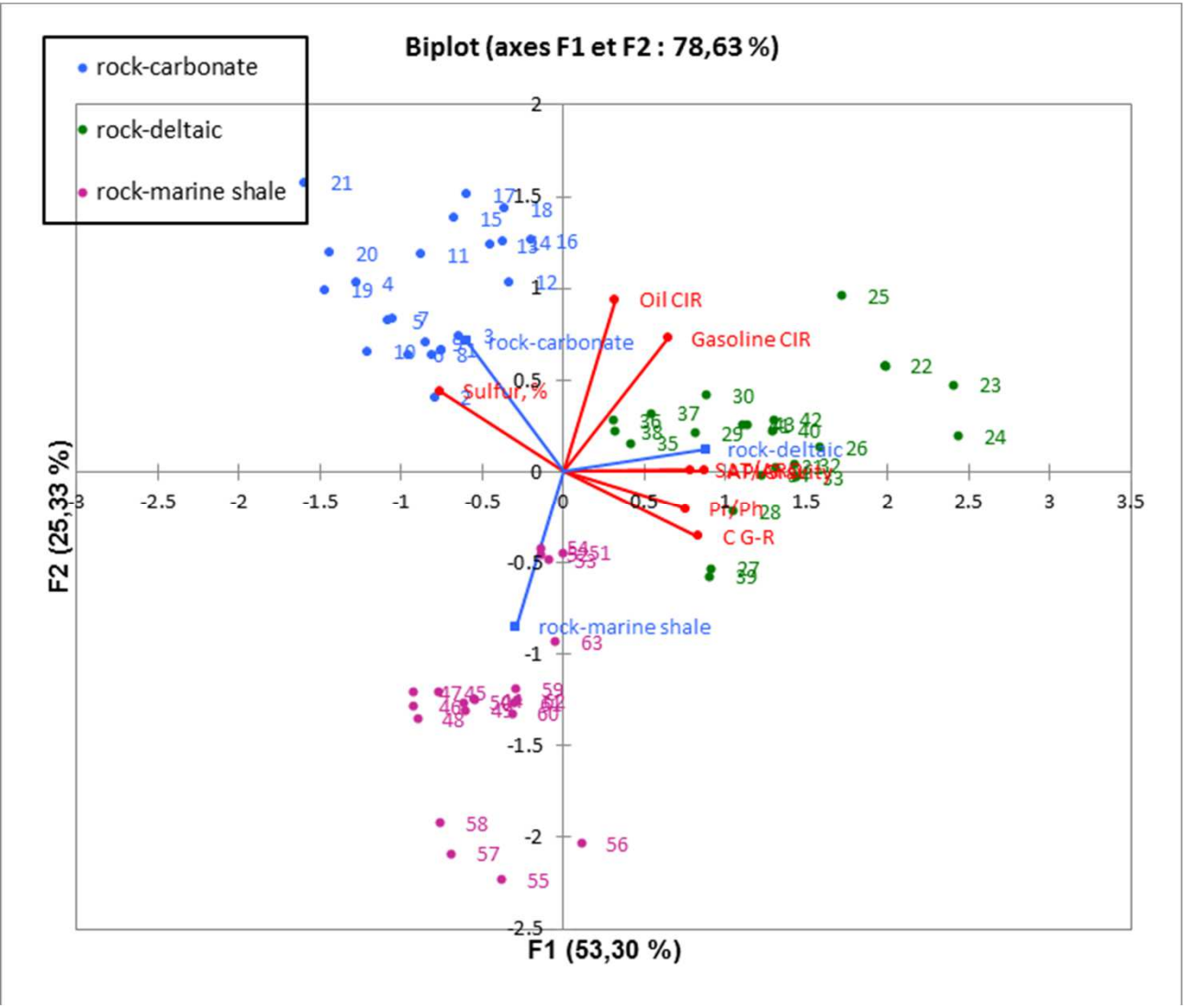


Regroupement des observations selon le type de roche
Observations « colorées »

5.4 représentations graphiques « biplots »



ACP : représentations graphiques « biplots »



ACP : conclusion

L’analyse en composantes principales :

- A pour principe un changement de base :
  - Représentation 2D de données multi-variées
  - La projection dans le nouvel espace limite les distorsions (modification des distances)

- Représente en priorité les directions où la variabilité est la plus importante
- Les nouvelles variables ne sont plus corrélées

ACP : conclusion

L’analyse en composantes principales :

- S’effectue sur un tableau de données de grandes dimensions (grand nombre d’individus)
- Prend en compte des variables quantitatives, discrètes ou continues
  - Le jeu de données est souvent centré et réduit avant l’ACP
- Peut inclure des données qualitatives en « coloration »
- Permet d’identifier des groupes d’individus ou de variables ayant un comportement similaire
- Peut s’utiliser en pré-traitement d’un jeu de données :
  - Permet d’identifier des valeurs aberrantes ou « outlier »
  - Pour déterminer un nombre de classes pertinent pour un futur classement

ACP : conclusion

L’interprétation d’une ACP se fait sur la base :

- Du tableau de corrélation : lien entre les variables

- Du « scree plot » : choix du nombre de variables latentes

- Des graphiques :
  - Cercle de corrélation entre les variables (« loading plot »)
  - Graphique des individus (« score plot »)
  - Graphique combiné variables + individus : « biplot »